http://pubsonline.informs.org/journal/deca/

DECISION ANALYSIS
Articles in Advance, pp. 1–24
ISSN 1545-8490 (print), ISSN 1545-8504 (online)

Identifying Soccer Players on Facebook Through Predictive Analytics

Matthias Bogaert,[a] Michel Ballings,[b] Martijn Hosten,[a] Dirk Van den Poel[a]

[a] Department of Marketing, Ghent University, 9000 Ghent, Belgium; [b] Department of Business Analytics and Statistics, University of Tennessee, Knoxville, Tennessee 37996

Contact: matthias.bogaert@ugent.be, http://orcid.org/0000-0002-4502-0764 (MaB); michel.ballings@utk.edu, http://orcid.org/0000-0002-3559-3112 (MiB); martijn.hosten@ugent.be (MH); dirk.vandenpoel@ugent.be, http://orcid.org/0000-0002-8676-8103 (DVdP)

Received: October 13, 2016
Revised: February 1, 2017; March 31, 2017; May 31, 2017
Accepted: June 7, 2017
Published Online in Articles in Advance:

https://doi.org/10.1287/deca.2017.0354

Abstract. This study assesses the feasibility of identifying self-reported sports practitioners (soccer players) on Facebook. The main goal is to develop a system to support marketers with the decision as to which prospects to target for advertising purposes. To do so, we benchmark several algorithms (i.e., random forest, logistic regression, adaboost, rotation forest, neural networks, and kernel factory) using five times twofold cross-validation. To evaluate performance and variable importances, we build a fusion model, which combines the results of the other algorithms using the weighted average. This technique is also referred to as information-fusion sensitivity analysis. The results reveal that Facebook data provide a viable basis to come up with sports predictions as the predictive performance ranges from 72.01% to 80.43% for area under the receiver operating characteristic curve (AUC), from 81.96% to 83.95% for accuracy, and from 2.41 to 3.06 for top-decile lift. Our benchmark study indicates that stochastic adaboost, the fusion model, random forest, rotation forest, and regularized logistic regression are the best-performing algorithms. Furthermore, the results show that the most important variables are the *average number of friends that play soccer, membership of a soccer group*, and the *number of favorite teams*. We also assess the impact of our results on profitability by conducting a thorough sensitivity analysis. Our analysis reveals that our approach can be beneficial for a wide range of companies. The analysis and results in this study will assist sports brands with decisions regarding their implementation of targeted marketing approaches.

Keywords: Facebook • sports • predictive models • social media • sensitivity analysis

1. Introduction

The sports industry is flourishing with opportunities in sports marketing, sponsorship, and marketing. Concurrently, competition to secure some of these opportunities has become more fierce (Belzer 2016). Hence, to survive and sell their products, sports brands need to adopt new ways to market their products, such as social media marketing, and embrace targeted advertising. A targeted advertising strategy implies that sports companies identify customers who are most likely to buy to subsequently send them advertisements (Burez and Van den Poel 2007). For example, to sell soccer gear, a viable strategy may be to identify and target soccer players. Before the existence of social media, companies had to rely on expensive commercial databases to implement targeted marketing strategy companies (D'Haen et al. 2016). However, the Internet and social media offer an inexpensive and effective alternative (Ballings and Van den Poel 2015a). In that regard, Facebook can be considered as the most important social media channel given that (1) it has 1.79 billion monthly active users (Facebook 2017) and (2) it contains a large number of variables related to customer engagement (e.g., likes and comments) (Malthouse et al. 2013). For example, users who are active in a soccer group on Facebook are more likely to be interested in soccer and thus might be more inclined to buy soccer gear.

Despite the fact that sports companies acknowledge social media plays an important strategic and operational role, advertising research in the sports industry has not yet advanced to using social media data in a predictive context (Filo et al. 2015). In the sense of targeted marketing approaches, to the best of our knowledge,

no study has looked into whether one can accurately predict which Facebook users are sports practitioners. This is an important question because if the answer is "yes" it would mean that the sports industry now has a large amount of data and contact information of 1.79 billion consumers at its disposal, as such solving the biggest data-sourcing problem that companies are currently facing. Secondary questions are the following: (1) Which variables are the driving force of these predictions? and (2) How are these variables related to the probability of being a sports practitioner? The answers to these questions will pave the way for companies, or Facebook, Inc.[1], for that matter, to implement a targeting system efficiently and effectively. For example, Facebook, Inc., could implement our targeted advertising approach to provide more customized targeting options to marketers.

To fill this gap in literature, this paper mines all available data in Facebook to evaluate the feasibility of identifying self-reported soccer practitioners (i.e., Facebook users who may be interested in buying soccer gear). It is important to note that we are not targeting all soccer players who are on Facebook, but instead all soccer players on Facebook who have self-reported to be a soccer player. Stated differently, we are targeting a subset of soccer players on Facebook. Our model can be used to make predictions for all users. For example, soccer players who did not self-report to be soccer players and have profiles similar to self-reported soccer players will receive a high predicted probability of being soccer players. We chose soccer as our focal sport because the soccer market is considered one of the biggest industries in Europe, and in effect the biggest sport, with revenues of the top 20 clubs rising up to $6.6 billion (Deloitte 2016). To that end, our Facebook data were collected via an application on the Facebook page of a European soccer team. This application was advertised several times on their Facebook page. To increase awareness and interest, an incentive (i.e., a signed jersey) was offered to use the application. Users who signed in on the application were informed that their Facebook data would be collected for academic purposes. Only data from users who gave their authorization were collected. As a consequence, our data set contains a lot of information specifically related to soccer. Next to assessing the capacity of Facebook data to accurately predict which Facebook users play soccer, we also contribute to the decision analysis process in the following ways. First, we benchmark several algorithms (i.e., logistic regression, neural network, rotation forest, random forest, stochastic adaboost, and kernel factory) to determine which algorithms work best on this problem. Second, we use information fusion to build a fusion model and determine which variables are important (Sevim et al. 2014). Information fusion is a technique that intelligently combines the results of different algorithms (Oztekin et al. 2013). To combine these results, information fusion takes the weighted averages of the results of the individual prediction models. This implies that models with a better performance will have a higher weight in the final prediction and variable importance score. Finally, we assess how our results can be used in practice by decision makers and evaluate the impact on firm profitability by conducting sensitivity analysis.

The remainder of this article is organized as follows. In Section 2 we provide an extensive literature review and highlight our contribution. In Section 3, we explain our methodology. Section 4 contains a discussion of the results. Section 5 elaborates on the implications of our results for decision makers. Section 6 summarizes the main conclusions. Finally, we discuss the limitations and avenues for future research in Section 7.

2. Literature Overview

Predicting whether or not someone is a soccer player falls under the most commonly used segmentation framework in marketing called AOI, short for activities, opinions, and interests (Hoch 1988). To highlight our contribution, we therefore review predictive studies in social media about AOI that can be used for a targeted advertising strategy. These predictive studies can be categorized according to the social media data they use and which activities, opinions, or interests they predict. Of all social media sites, Facebook and Twitter are most widely adopted (Shapira et al. 2012, Tumasjan et al. 2010). Other social media platforms that are used for predicting AOI are Instagram (Cesario et al. 2016), Flickr (Kisilevich et al. 2010), and Foursquare (Gao et al. 2013). Predictions of "interests" on social media have already been applied in several fields, such as music (Passant and Raimond 2008), movies (Quijano-Sanchez et al. 2011), items (Guy et al. 2010), and online communities (Baatarjav et al. 2008). For example,

Table 1. Overview of Social Media Literature for Targeting

Author	Data	Places	Music	Movie	Item	Group	Event	User mobility	Political preferences	Sports
Berjani and Strufe (2011)	Gowalla	X								
Chang and Sun (2011)	Facebook	X								
Gao et al. (2013)	Foursquare	X								
Passant and Raimond (2008)	MySpace/Last.fm		X							
Bu et al. (2010)	Last.fm		X							
Mesnage et al. (2011)	Facebook		X							
Quijano-Sanchez et al. (2011)	Facebook			X						
Shapira et al. (2012)	Facebook			X						
Said et al. (2011)	MoviePilot			X						
Guy et al. (2010)	Lotus				X					
Baatarjav et al. (2008)	Facebook					X				
Kim and Saddik (2013)	Last.fm					X				
Rosaci and Sarne (2014)	Facebook					X				
Carmagnola et al. (2009)	Facebook					X				
Kayaalp et al. (2009)	Last.fm						X			
Zhang et al. (2013)	Facebook						X			
Bogaert et al. (2016a)	Facebook						X			
Cesario et al. (2015)	Twitter							X		
Cesario et al. (2016)	Instagram							X		
Kisilevich et al. (2010)	Flickr							X		
Tumasjan et al. (2010)	Twitter								X	
Golbeck and Hansen (2011)	Twitter								X	
Our study	Facebook									X

Shapira et al. (2012) combine profile data with friends data to come up with reliable movie recommendations for a certain user. Regarding "opinion" prediction studies, several authors have tried to accurately predict political orientation (Golbeck and Hansen 2011) and election outcomes (Tumasjan et al. 2010) using Twitter data. Finally, activities on social media have already been studied in several fields of application. For example, check-in behavior (Chang and Sun 2011), event attendance (Bogaert et al. 2016a), and movement patterns (Cesario et al. 2015) have been studied using different social media platforms. Table 1 provides a representative overview of the most important AOI studies that can be used for targeted advertising.

From Table 1, it is clear that there are no studies that investigate how Facebook data can be used to make accurate sports predictions. In more general terms, we can state that no study has investigated the feasibility of predicting someone's hobbies. Table 1 shows that certain activities, such as places, events, and user mobility, already have been studied. However, the question whether or not a person's leisure activities can be predicted using social media data is, to the best of our knowledge, not yet researched. The capacity to predict who is more likely to practice a certain sport (or, more in general, a hobby) opens a lot of targeted advertising opportunities for decision makers (e.g., sport retailers and sport brands). For example, if sports companies want to acquire new customers, the standard approach is to conduct market research to segment and profile the customer base. In a later phase, the company then tries to target these customers through Internet, TV, or radio advertising. A problem with this approach is that the reach of those advertisements is often too wide, resulting in extra costs for the company (e.g., people of other segments click on a web advertisement for which the company also pays per click). Another option is to come up with a one-to-one targeting approach (Burez and Van den Poel 2007). Whereas before companies had to rely on expensive internal company databases to implement such one-to-one strategies, the Internet and, in particular, social media websites, now offer a viable alternative (D'Haen et al. 2016). The question of whether or not Facebook data is able to drive such a targeting approach in sports or, in general, leisure activities is therefore of major importance in decision analysis. It allows decision makers to answer questions such as the following: (i) "Which customer do we need to

target?" (ii) "Where can we find prospects with a high likelihood of becoming a customer?" (iii) "Which cost-effective data source can we use to drive accurate and reliable results?" In that regard, Facebook has considerable potential to be a cost-effective data source. First, Facebook has a very large amount of data in terms of user behavior, characteristics, and preferences (Lampe et al. 2007). Second, Facebook is the fastest growing social media platform with 1.79 billion monthly active users (24% of the world population) (Facebook 2017). In addition, Facebook is considered to be the most effective advertising channel of all social media platforms (Egan 2016).

This study builds on several areas of literature. First, we aim to fill the gap in extant literature by assessing the feasibility of accurately predicting which Facebook users play soccer (we do not distinguish between professional, semiprofessional, or nonprofessional players). In general terms, we try to assess the capacity of Facebook to accurately predict a given user's hobbies and leisure activities. Given that there exists a large number of application domains, and the predictive potential in each domain is different, we feel that this is a valuable contribution. This study adds a piece to the puzzle in AOI prediction. By doing so, we contribute to the generalizability of Facebook as a data source for one-to-one targeting. To make these predictions, we have gathered data via a Facebook application that we developed for a European soccer team. Hence, we have many soccer-related variables and soccer players in our database. Previous research has shown that including Facebook data contains many valuable variables and can lead to accurate recommendations (Kalampokis et al. 2013). For example, Ballings and Van den Poel (2015a) revealed that accurate predictions can be achieved for Facebook usage frequency using a wide variety of user-related variables. Bogaert et al. (2016a) went one step further and showed that the inclusion of Facebook friends is beneficial in event prediction.

The second contribution of this study is that we test different algorithms to determine which one has the best predictive performance in the field of sport predictions. In total, we benchmark six algorithms: logistic regression (Guisan et al. 2002), neural network (Baldi and Hornik 1989), rotation forest (Rodriguez et al. 2006), random forest (Breiman 2001), stochastic adaboost (Friedman 2002), and kernel factory (Ballings and Van den Poel 2013). In addition, we also include a fusion model, which can be seen as an ensemble of all other prediction models (Oztekin et al. 2016).

Third, we use variable importance measures to determine which variables are most important. In line with previous research on Facebook data, we can make hypotheses about which variables will be important. It has been shown that Facebook likes (Passant and Raimond 2008), user characteristics and preferences (Zhang et al. 2013, Hsu et al. 2012), and social network variables (Zhang et al. 2013, Konstas et al. 2009) are important. Moreover, a lot of these variables are related to fan engagement and loyalty (Funk 1998, Yoshida et al. 2014, Bauer et al. 2008). For example, users can indicate their favorite team on Facebook and like pages of sports teams. Thus, if a user has a favorite team or a certain sport and likes a lot of pages related to one sport, this can serve as an indicator of high sport engagement, as such increasing the propensity of playing a certain sport. Moreover, research on fan engagement has shown that highly engaged customers are more loyal and are also more likely to buy sports gear (Bauer et al. 2008). As such, we can hypothesize that the *number of sports that a user practices* and variables related to likes (e.g., *number of favorite teams*, *number of sports pages liked*) will be among the most important variables.

3. Methodology

3.1. Data

To gather data from profiles of Facebook users, we developed a Facebook application. This application has been built for a European soccer team. To stimulate usage of our application, we performed several actions. First, we offered a signed football shirt of a well-known soccer player as a prize to the person who could correctly answer several questions about the soccer club. Second, we regularly advertised our application on the Facebook page of the European soccer team as this page is very often visited since it is one of the most famous teams in Europe. Third, to increase awareness, we also added the application to the main page tabs. To avoid privacy issues, users of the Facebook application were immediately presented an authorization box where they had the possibility to give permission to the application to extract data from their Facebook user profile. Along with this authorization

box, we added a section with rules and regulations, which also included our contact information. In addition, we promised that all information would be anonymous and that we would not extract private messages. Finally, we also added information about the purpose of our research. The authorization box and data extraction were followed by some questions to determine the winner of the prize. The data collection started on May 7, 2014, and finished on June 9, 2014. Our data consists of 5,010 unique observations, and every observation represents a unique Facebook user. Next to user data, we also included information about the Facebook friends of our participants serving as network variables.

The dependent variable in our study is binary {0: non-soccer player, 1: soccer player}. On Facebook, users can indicate which sports they practice. Users who entered soccer were considered to be soccer players. For the entire collection of observations, 15.44% of the Facebook participants play soccer, and 84.66% do not. To cope with class imbalance, we oversampled the response variable (Bogaert et al. 2016b).

3.2. Variables

We have included different types of independent variables that describe the user's characteristics, preferences, and Facebook behavior. The following variables describe the characteristics of the user: demographic and identification variables (e.g., gender and age), professional and educational variables (e.g., work and schools), geographical variables (e.g., location and hometown), social variables (e.g., groups and relationship status), and general Facebook account variables (e.g., length of relationship). The preference variables of the user include a user's likes concerning, for example, religion, politics, books, movies, and music. Finally, Facebook behavioral variables encompass the different posts of a user (e.g., statuses, photos, videos, and albums) and the user's interactions with other users (e.g., comments made and received, likes and tags).

The aforementioned categories of variables can be related to the Facebook user specifically and are, in that case, included as Facebook user variables. For example, a certain user is a male, who lives in New York, working in finance, is in a relationship, likes soccer, and has posted 15 photos and made 17 comments on statuses. These variables can also refer to the user's friends. In that case, they are calculated as the average value over all friends of a given user. We note that we only included a selection of the Facebook friends variables. More specifically, we only added variables that are related to sports and soccer. For example, the *average number of soccer-related likes and groups* of Facebook friends are included while the *average number of status updates over all friends* are excluded.

Table 2. Summary of Sports-Related Variables (Other Categories Are Not Shown Because of Space Constraints)

Variable
COUNT(*sports other than soccer*)
IND(*soccer groups*)
COUNT(*favorite teams*)
COUNT(*books/music/television programs/check-ins related to sport/soccer*)
COUNT(*likes related to sport/soccer*)
COUNT(*sport/soccer events*)
COUNT(*interest == sport/soccer*)
AVERAGE(*friends playing soccer*)
AVERAGE(*soccer groups over all friends*)
AVERAGE(*likes related to sport/soccer over all friends*)
AVERAGE(*sport/soccer events over all friends*)
AVERAGE(*books/music/television programs/check-ins related to sport/soccer over all friends*)
AVERAGE(*interest == sport/soccer over all friends*)
AVERAGE(*number of sports over all friends*)

The like variables in our study are variables related to likes generated by the user specifically and thus not likes generated by a Facebook friend when, for example, liking the user's post. These likes are only available for a page (e.g., soccer team), music group, and leisure activity. Since our dependent variable denotes whether or not someone is a soccer player, we summarize all variables related to the sports category in Table 2. Because of space constraints, we do not show variables unrelated to soccer as these are likely to be less important than soccer-related variables. *COUNT* refers to the frequency. For example, the variable *COUNT*(*sports other than soccer*) sums up all the sports a user practices besides soccer. *AVERAGE* stands for the average frequency over all friends (e.g., *average number of soccer groups a user's friends are part of*). *IND* refers to an indicator variable (e.g., *whether or not the user is part of a soccer-related group*). In total, we have 520 variables in our model of which 56 are friend-related variables and 464 are user-related. All variables are either Boolean, numeric, or integer.

3.3. Classification Algorithms

In this section, we explain the classification algorithms used in this study. We use several single classifiers and ensemble methods. The single classifiers include logistic regression (LR) and neural network (NN). The ensemble methods are random forest (RF), adaboost (AB), kernel factory (KF), and rotation forest (RoF). For a detailed description of the classification techniques, we refer the reader to Appendix A.

3.3.1. Regularized Logistic Regression. Logistic regression uses the logistic (or sigmoid) function to link a binary dependent variable and a set of independent variables. To fit the model, logistic regression uses maximum likelihood. The predicted probability scores are restricted to $[0, 1]$ (James et al. 2013, p. 130). A common issue with logistic regression is overfitting (i.e., the model estimates random error or noise instead of the underlying relationship) (Babyak 2004). To avoid the problem of overfitting, we use the lasso approach to regularized logistic regression. Lasso, which stands for least absolute shrinkage and selection operator, imposes a bound on the total sum of the absolute values of all coefficients. Hence, it shrinks the coefficients toward zero (Guisan et al. 2002, James et al. 2013).

3.3.2. Neural Networks. Neural networks is a nonlinear classification technique that fits models by mimicking the behavior of the human brain. A neural network consists of three layers, namely the input layer, one or more hidden layers, and the output layer (Baldi and Hornik 1989, Dreiseitl and Ohno-Machado 2002). The input layer represents the independent variables, and the output layer represents the dependent variable. The hidden layer handles complexity with the activation function (e.g., sigmoid function; Specht 1990). We use artificial neural networks optimized by the BFGS algorithm with one hidden layer of neurons. This approach has good performance in terms of efficiency and reliability (Ballings et al. 2015, Dreiseitl and Ohno-Machado 2002).

3.3.3. Random Forest. Random forest is an ensemble learning method that uses bagging in combination with random feature selection (Dudoit et al. 2002). This means that the random forest algorithm trains every tree on an independently bootstrapped sample of the initial data set while each node split is determined by searching across a randomly selected subset of variables (Breiman 2001). To construct the ensemble, every tree votes for the most popular class, and the final prediction is determined by rule of majority voting (Breiman 2001). By doing so, random forest decorrelates the trees and lowers the variance of the classification errors (Chan and Paelinckx 2008, Gislason et al. 2006).

3.3.4. Adaboost. Boosting is a general ensemble method for improving the performance of algorithms. In adaboost, every model is built in a sequential way by reweighting the training data (James et al. 2013, p. 322). As such, each model is dependent on the previous one. Every misclassification in the current model gets a higher weight in the next iteration, and a correctly classified observation gets a lower weight (Chan and Paelinckx 2008, Freund et al. 1996). Adaboost thus gives a higher importance to the observations that are hard to classify (Chan and Paelinckx 2008, Freund et al. 1996). In contrast to other ensemble methods, adaboost lowers both the bias and variance component of the classification errors (Friedman 2002).

3.3.5. Kernel Factory. Ballings and Van den Poel (2013) propose an ensemble method for kernel machines. By using a row and column parameter, they create a number of mutually exclusive partitions by randomly splitting the training data followed by scaling. They apply the *burn* method to automatically select the best kernel function to transform each partition. They use each partition as training data for a single base classifier (e.g., random forest). Finally, they combine the number of predictions equal to the number of partitions into one final prediction by using a weighted average. They determine the optimal weights by applying a genetic algorithm.

3.3.6. Rotation Forest. Rotation forest is a classifier ensemble technique-based feature extraction (Rodriguez et al. 2006). The training set for each base classifier (i.e., a decision tree) is formed by randomly partitioning the feature set into K disjoint subsets and applying principal components analysis (PCA) to each subset (De Bock and Van den Poel 2011). By doing so, a rotation of the feature axes takes place without reducing the variability in the information of the data. Rotation forest promotes diversity by applying PCA and accuracy by keeping all principal components and using the whole data set as training data (Rodriguez et al. 2006).

3.4. Performance Evaluation

To evaluate our models, we use the most commonly used performance measures in CRM: accuracy, the area under the receiver operating characteristic curve (AUROC or AUC), and top-decile lift (Coussement et al. 2010). The accuracy or the percentage correctly classified is defined as follows (He and Garcia 2009):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{1}$$

with TP (true positives), FN (false negatives), FP (false positives), TN (true negatives).

We work with probabilistic output since we are interested in ranking users by their likelihood of playing soccer. To determine which instance is considered to be a positive or negative example, we calculate the accuracy with rate-fixed threshold (Hernandez-Orallo et al. 2012). A rate-fixed threshold implies that a certain percentage of the sample needs to have a score above a given threshold (Hernandez-Orallo et al. 2012). Hence, in our case, we calculated the accuracy after dichotomizing the propensity scores using a threshold that results in the top 10% of users being assigned a 1 and the remaining users being a 0. This model evaluation corresponds to the real-life situation in which 10% of the users will be targeted (Ballings and Van den Poel 2015a).

The downsides of working with a rate-fixed threshold are that (1) it is dependent upon the distribution of the data (i.e., sensitive to high class imbalance) and (2) it only considers one specific cutoff (i.e., highly dependent upon the chosen threshold) (He and Garcia 2009). To overcome these problems, we use the most commonly used portmanteau measure: the AUC. The AUC—and portmanteau measures in general—uniformly aggregate over all possible rates and assign an equal weight to all classes (Hernandez-Orallo et al. 2012). As a consequence, the AUC is an appropriate measure when the operating conditions are unknown and the data are unbalanced. The receiver operating characteristic curve (ROC curve) is a graphical depiction between the sensitivity and one minus specificity for all possible cutoff values (Ulvila and Gaffney 2004). Sensitivity is also called the detection rate or true positive rate (i.e., the ability of a model to identify soccer players), and 1 minus the specificity is often referred to as the false positive rate (i.e., the proportion of false alarms). In that case, the AUC can be seen as the probability of identifying a soccer player given a certain false positive rate (Ulvila and Gaffney 2004). Sensitivity and specificity are defined as follows (James et al. 2013, p. 148):

$$Sensitivity = \frac{TP}{TP + FN}, \tag{2}$$

$$Specificity = \frac{TN}{TN + FP}, \tag{3}$$

and AUC calculates the area under the ROC curve, which is defined as (Hanley and McNeil 1982)

$$\text{AUC} = \int_0^1 \frac{TP}{TP + FN}\, d\frac{FP}{FP + TN} = \int_0^1 \frac{TP}{P}\, d\frac{FP}{N}, \tag{4}$$

with $\boldsymbol{P}$ (positives) and $\boldsymbol{N}$ (negatives).

The AUC value can be interpreted as the probability that a randomly chosen positive example is ranked higher than a randomly chosen negative example (Hanley and McNeil 1982). Hence, it displays the ability of a classifier to avoid misclassification. The AUC can take values ranging from 0.5 to 1. A value of 0.5 indicates that the predictions are not better than predictions by a random model whereas a value of 1 indicates that the predictions are perfect (Langley 2000, Ulvila and Gaffney 2004). In contrast to the accuracy, AUC remains unchanged across different cutoff values. More specifically, accuracy only considers one specific cutoff value whereas AUC includes the entire range of possible cutoff values (Ballings and Van den Poel 2012, Langley 2000).

Next to AUC, another measure often used in CRM applications is the top-decile lift (Burez and Van den Poel 2007). The top-decile lift only focuses on the prospects with the highest probability of converting (i.e., the top 10% proportion of users) (Coussement et al. 2010). The top-decile lift is defined as the proportion of identified customers in the top 10% compared with the proportion of customers in the total data set (Lemmens and Croux 2006):

$$Lift = \frac{P_{\text{top }10\%}/(P_{\text{top }10\%} + N_{\text{top }10\%})}{P/(P + N)}. \tag{5}$$

Hence, the top-decile lift evaluates how much better our model is in detecting soccer players in the top 10% compared with a random model. This implies that a higher top-decile lift denotes a better model (De Bock and Van den Poel 2011). A random model has

a top-decile lift of 1; a perfect model has a top-decile lift equal to one divided by the proportion of positive cases in the data set (in our case, $1/0.154 = 6.4935$). It is clear that the top-decile lift is an interesting measure for decision makers since it implicitly assumes that marketing budgets are limited and only a certain percentage of the users can be targeted (Coussement et al. 2010). Also, the top-decile lift is directly related to profitability and thereby allows decision makers to calculate the impact of their prospecting model on firm performance (Lemmens and Croux 2006).

3.5. Cross-Validation

To obtain results that are not overly optimistic or pessimistic, we use five times twofold cross-validation (5×2cv). This process starts by randomly dividing the data in two samples. Each sample is used once as a test sample and once as a training sample. This process is repeated five times. As such, this results in 10 AUCs, accuracies, and top-decile lifts per model (Dietterich 1998). We take the median of all AUCs (accuracies and top-decile lifts) per model to end up with the overall AUC (accuracy and top-decile lift) for each model while using the interquartile range (IQR) as a measure of dispersion.

To test for significant difference between the performance of our classifiers, we use the Friedman test with Bonferroni–Dunn post hoc test as suggested by Demšar (2006). The Friedman test compares the average ranks of the algorithms to find significant differences. To evaluate which algorithms differ significantly, we use the Bonferroni–Dunn test. The performance of two classifiers differs significantly when the corresponding ranks differ with at least the critical difference (in our case, 2.548799). For a more detailed explanation, we refer to Demšar (2006).

3.6. Information-Fusion Sensitivity Analysis

Researchers commonly agree that there exists no single best technique that works for every data set and for every application. Therefore researchers often aggregate results of several algorithms to obtain more accurate and precise results (Predd et al. 2008, Wang et al. 2011). Dietterich (2000) states that ensembles solve the representation, statistical, and computational problem of single classifiers. In that mindset, information fusion intelligently combines the results of different classifiers to extract more useful and accurate information in comparison to single classifiers (Oztekin et al. 2013). Hence, instead of using the results of one prediction model, information fusion combines all the available information of all prediction models (Oztekin et al. 2016). Given a dependent variable y and a set of independent variables X with $X = \{x_1, x_2, \ldots, x_n\}$, a classifier i can then be represented as

$$\hat{y}_{\text{individual}_i} = f_i(x_1, x_2, \ldots, x_n) = f_i(X). \tag{6}$$

The classifier f_i can take on many forms. For example, in the case of regression, our classifier f_i becomes the following:

$$f_i(X) = \beta + AX^T. \tag{7}$$

In Equation (7), β represents the intercept, and A are the coefficients of X with $A = \{a_1, a_2, \ldots, a_n\}$. Given that we have k classifiers, information fusion is defined as follows (Sevim et al. 2014):

$$\begin{aligned}\hat{y}_{\text{fusion}} &= \Psi(\hat{y}_{\text{individual}_1}, \hat{y}_{\text{individual}_2}, \ldots, \hat{y}_{\text{individual}_k}) \\ &= \Psi(f_1(X), f_2(X), \ldots, f_k(X)).\end{aligned} \tag{8}$$

In Equation (8), Ψ is the fusion operation. If we assume the Ψ is a linear combination of classifiers f_i with α_i as the individual weighting coefficient of each classier f_i, then Equation (8) can be reformulated as

$$\hat{y}_{\text{fusion}} = \sum_{i=1}^{k} \alpha_i f_i(X) = \alpha_1 f_1(X) + \alpha_2 f_2(X) + \cdots + \alpha_k f_k(X), \quad \text{where } \sum_{i=1}^{k} \alpha_i = 1. \tag{9}$$

The values of α are the weighted average of the predictive performance of each classifier $f_i(x)$. Hence the better the predictive performance of the individual prediction model, the higher their α values and the larger their weight in the fusion function Ψ (Oztekin et al. 2013). This implies that information extracted from highly accurate models will receive higher weight than poorly performing algorithms. In our case, we use the weighted average of the 5×2cv median AUCs for our α values.

Next to determining our fusion function, another important aspect of decision analysis is to assess which variables are the driving force of predictive performance (Sevim et al. 2014). The most common way of conducting sensitivity analysis in data mining is by means of variable importances (VIM). VIM capture the

effect on predictive performance of permuting on a certain variable. Hence VIM can be seen as a form of sensitivity analysis since they show how the output varies as we change the input values (Wei et al. 2015). The higher the change in predictive performance when permuting on a certain variable, the higher its sensitivity and the higher its variable importance. Several techniques for calculating VIM have been proposed in literature, such as the mean decrease in Gini index or mean decrease in accuracy (Breiman 2001). However, a problem with most of these techniques is that the underlying performance measures are sensitive to the underlying distribution of the data (Janitza et al. 2013). To alleviate this problem, we use the mean decrease in AUC as proposed by Janitza et al. (2013). The mean decrease in AUC is more robust to changes in the distribution since it uses the AUC to determine the change in predictive performance. If we then rewrite Equation (9) in terms of an information fusion–based sensitivity measure of the variable n with k prediction models (i.e., mean decrease in AUC after permuting on the variable n), we obtain Equation (10):

$$V_{n(\text{fusion})} = \sum_{i=1}^{k} \alpha_i V_{in} = \alpha_1 V_{1n} + \alpha_2 V_{2n} + \cdots + \alpha_k V_{kn}. \quad (10)$$

In Equation (10) V_{in} stands for the variable importance measure of variable n in prediction model i. The values of α are the same as in Equation (9), namely the weighted average of the median 5×2cv AUCs of the different classifiers.

4. Results

4.1. Model Performance

All performance results were obtained by running the algorithms on a server with 128 GB RAM and 24 cores operating at 2.67 GHz. The run time of all algorithms combined was approximately 10 hours. Some algorithms (random forest, logistic regression, and rotation forest) had run times of a few minutes. Other algorithms (adaboost, neural networks, and kernel factory) took several hours. The reason for this discrepancy is that adaboost works sequentially, and neural networks and kernel factory have to perform an extensive grid search for the optimal parameter settings. Another reason is that some algorithms have a faster C++ or FORTRAN backend whereas other algorithms rely on R-code.

Table 3. 5×2cv Median AUC, Accuracy, and Lift

	LR	RF	AB	KF	NN	RoF	Fusion
AUC	0.7646	0.7875	0.8043	0.7510	0.7201	0.7761	0.8007
Accuracy	0.8295	0.8271	0.8395	0.8232	0.8196	0.8255	0.8323
Lift	2.7190	2.6410	3.0550	2.5115	2.4080	2.6020	2.8220

Table 3 provides an overview of the cross-validated results. We also included the fusion model based on Equation (8), which can be seen as an ensemble of the different algorithms. The model performance is calculated as the median AUC, accuracy, and top-decile lift for each algorithm. The main research question was the following: "Can we accurately predict, using Facebook data, if someone is a self-reported soccer player or not?" The results clearly indicate that predicting whether a user plays soccer is a viable approach since the median AUC ranges from 72.01% to 80.43% and the median accuracy from 81.96% to 83.95%. The top-decile lift shows us how much better our model is at identifying soccer players in the top 10% of the predictions as opposed to randomly selecting users. Table 3 shows that the best algorithm is 3.0550 times better at detecting soccer players than a random model. This measure offers an indication that the strategy we propose in this study affords decision makers to set up an effective targeting campaign.

Table 3 also indicates that the adaboost model is the top-performing algorithm across all performance measures. Kernel factory (KF) and neural networks (NN) come in last. In terms of AUC, the fusion model is the second best performer, followed by random forest (RF), rotation forest (RoF), and logistic regression (LR). However, in terms of accuracy and top-decile lift, the fusion model is followed by logistic regression. Hence, this implies that, over the whole range of cutoff values, random forest and rotation forest perform better than logistic regression, but when focusing on the top 10%, proportion logistic regression is superior.

Table 4 provides an overview of the average ranks and the Friedman test combined with the Bonferroni–Dunn post hoc test. The Friedman test statistic indicates that we can reject the null hypothesis of no significant differences between classifiers for AUC, accuracy, and top-decile lift. To determine which classifiers perform significantly worse than our top performer, we used the Bonferroni–Dunn post hoc test. In doing so, we can

Table 4. Average Ranks Based on AUC, Accuracy, and Lift

	LR	RF	AB	KF	NN	RoF	Fusion	Friedman χ^2 (6)
AUC	5	**3.10**	**1.30**	5.40	6.70	4.70	**1.80**	50.74, $p < 0.001$
Accuracy	**3.65**	4.65	**1.35**	5.35	6.35	4.55	**2.10**	40.63, $p < 0.001$
Lift	**3.85**	4.05	**1.30**	5.55	6.50	4.55	**2.20**	42.18, $p < 0.001$

Table 5. Interquartile Ranges

	LR	RF	AB	KF	NN	RoF	Fusion
AUC	0.0221	0.0068	0.0082	0.0104	0.0131	0.0896	0.0064
Accuracy	0.0058	0.0026	0.0044	0.0036	0.0062	0.0178	0.0086
Lift	0.2070	0.1170	0.1420	0.1355	0.1943	0.5568	0.2585

divide the classifiers into two groups based on their performance compared with the best-performing algorithm. Classifiers for which the difference in average rank is greater than the critical difference (2.548799) perform statistically worse than the best-performing algorithm; others have an equal performance in statistical terms. The latter are highlighted in bold in Table 4. More specifically, Table 4 indicates that adaboost, the fusion model, and random forest perform equally in terms of AUC. For accuracy and top-decile lift, logistic regression performs equally well as adaboost and the fusion model. This difference can be explained by the fact that AUC is an aggregate performance measure and hence considers all possible cutoffs whereas accuracy only considers a threshold corresponding to the selection of the top 10% of instances. Hence, random forest performs well across the whole range of cutoff values whereas logistic regression performs better when considering the top 10%.

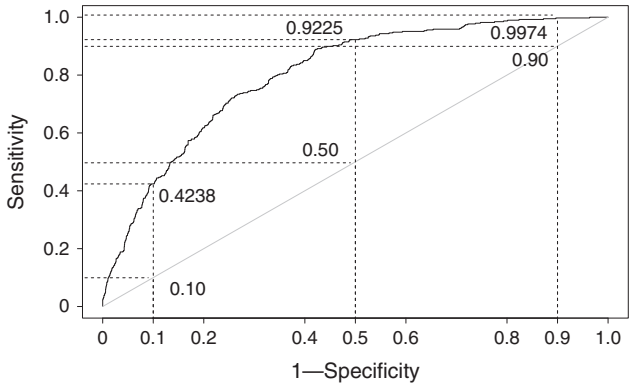
The viability of predicting whether someone plays soccer is also confirmed when looking at the stability of our results. As a measure of stability, we use the interquartile range (IQR). Table 5 displays the interquartile range per algorithm for AUC, accuracy, and top-decile lift. The IQR ranges from 0.64% to 8.96% for the AUC, from 0.36% to 1.78% for accuracy, and from 11.70% to 55.68% for top-decile lift. In terms of AUC, the fusion model achieves the most stable results. In terms of accuracy, rotation forest outperforms all others, and in terms of top-decile lift, random forest takes first place. These findings substantiate our hypothesis that Facebook data can be used to predict whether a Facebook user plays soccer or not.

4.2. ROC Analysis

To determine the impact of true positives and false negatives on classifier performance, we added a ROC analysis for our top-performing algorithm. We choose adaboost as our top-performing algorithm since it achieved the best results on two out of three performance measures and its performance is equal to the fusion model in statistical terms. Moreover, the run time of adaboost is much lower in comparison to the fusion model. The ROC curve denotes the trade-off between the true positive rate (TPR) and the false positive rate (FPR) (Ulvila and Gaffney 2004). The FPR informs decision makers of the cost of marketing to a few "non-soccer players" while the TPR is a linear function of missing true "soccer players" (i.e., false negative rate (FNR) or 1 – sensitivity). Hence, the ROC curve allows managers to make an informed trade-off between the benefits (i.e., revenues) of a marketing campaign and the costs of a marketing campaign. If a marketer wants to focus on generating revenue, regardless of the costs, the marketer will choose a point on the right-hand side of the ROC curve. If a marketer wants to minimize the costs of an advertising campaign, the marketer will choose a point on the left-hand side of the ROC curve. This implies that each point on the ROC curve refers to a specific propensity threshold, FPR and TPR, which, in turn, is related to a specific revenue–cost scenario (Ulvila and Gaffney 2004). Figure 1 plots the 5×2cv median ROC curve for adaboost (black solid line) and presents three possible scenarios. The straight grey solid line at 45 degrees in Figure 1 represents a random model.

In the first case, the marketer decides that the marketer wants to lower the costs of the advertising campaign. The marketer therefore decides to allow a FPR of 0.10, which corresponds to a TPR of 0.4238 for adaboost. In this case, the cost of the campaign will be low, but the revenues will also be lower. If the marketer would choose a random model (TPR = 0.10), revenues would be much lower. In the second case, the manager chooses a point in the middle of the ROC curve (FPR = 0.50). In this scenario, the cost of the campaign will be larger, but the campaign will target a lot of soccer players. Again, we see that there is a large difference between the performance of adaboost (TPR = 0.9225) and the random case (TPR = 0.50). In the final scenario, the manager chooses a point on the far right of the ROC

Figure 1. 5 × 2cv Median ROC Curve of Adaboost

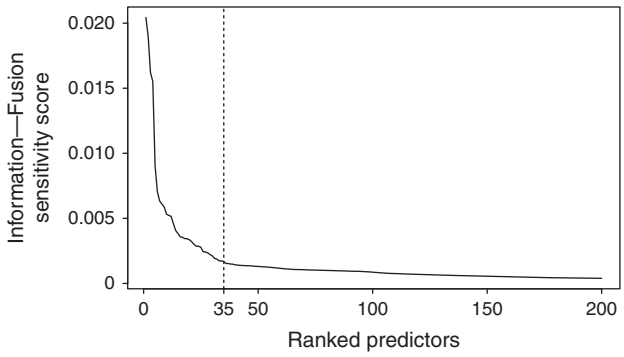


curve (FPR = 0.90 and TPR = 0.9974). In this case, the advertising strategy will target most of the soccer players, but the costs of targeting nonplayers will be considerably large. The difference between adaboost (TPR = 0.9974) and the random model (TPR = 0.90) decreases in this scenario. In the managerial implications (Section 5), we conduct a thorough sensitivity analysis on how these three scenarios affect profitability.

4.3. Information-Fusion Sensitivity Analysis

We made a scree plot (Figure 2) to evaluate which variables are most important in predicting whether a Facebook user plays soccer. The scree plot shows the top 200 variables in decreasing order by the 5 × 2cv information-fusion sensitivity score. The sensitivity score is calculated by means of Equation (10). The values of α are the weighted average 5 × 2cv median AUCs displayed in Table 3. The variable importances are calculated as the median 5 × 2cv mean decrease in AUC for each model. The final sensitivity score is then obtained by inputting the α values together with the 5 × 2cv variable importances into Equation (10). This technique is preferred above the traditional VIM since it integrates the results of all predictions into one measure of sensitivity (Oztekin et al. 2013). Also, in contrast to the traditional VIM, our sensitivity score explicitly incorporates the predictive performance of our prediction models. Figure 2 then orders the sensitivity scores from high to low and plots the scores against their respective ranks. From Figure 2, it is clear that variables with a rank higher than 35 only add little to the predictive performance. Because of space constraints, we only summarize the top 15 variables (Table 6). We refer the reader to Appendix B for the top 35 variables.

Figure 2. Scree Plot of the Sensitivity Score of All Variables



We can divide the most important variables into three major categories. In Table 6, *S* represents sports variables, *G* general Facebook variables, and *N* network variables. First, 66% of the top variables are general Facebook variables. For example, the *gender of the user* is one the most important variables. If the user is male, the prospensity of being a self-reported soccer player is larger. Second, 27% of the variables are related to sports, such as *number of sports* and *number of favorite teams*. The importance of these variables can be explained by the fact that they all indicate a certain degree of fan loyalty and engagement (Bauer et al. 2008). The higher the degree of engagement, the more positive the attitude toward soccer and the higher the chances of playing soccer (Funk 1998). Third, 7% of the variables are network variables (e.g., the *average number of friends who play soccer*). The *average number of friends who play soccer* is the most important variable.

Table 6. Top 15 Variables Based on Information Fusion

Rank	Variable name	Sensitivity score	Type
1	*AVERAGE(friends playing soccer)*	0.0204	*N*
2	*IND(soccer group)*	0.0189	*S*
3	*COUNT(favorite teams)*	0.0162	*S*
4	*IND(gender == female)*	0.0155	*G*
5	*IND("interested in" present)*	0.0090	*G*
6	*IND(relationship status == single)*	0.0070	*G*
7	*Age*	0.0063	*G*
8	*COUNT(likes related to soccer)*	0.0061	*S*
9	*IND(biography present)*	0.0059	*G*
10	*IND(education == college)*	0.0053	*G*
11	*COUNT(likes related to sport)*	0.0052	*S*
12	*COUNT(television shows)*	0.0052	*G*
13	*IND(relationship status == relationship)*	0.0046	*G*
14	*IND(relationship status == married)*	0.0041	*G*
15	*COUNT(friends)*	0.0038	*G*

Hence, the more friends who play soccer, the higher the propensity that the focal user is a soccer player as well. This finding is in line with Bogaert et al. (2016a), who found, based on homophily (McPherson et al. 2001), that the number of friends who attend an event was amongst the top predictors of event attendance.

In sum, we found that the most important variables can be grouped into different types. First of all, we showed that several important variables are directly related to sport and fan engagement (e.g., *number of favorite teams*, *membership of a soccer group*, and *number of sport likes*). Second, social network variables, such as the *number of friends playing soccer*, contribute strongly to our predictions. Third, several general Facebook variables have a strong impact on our dependent variable. In Appendix C, we explain the relationship between the most important variables and the response more in depth using partial dependence plots.

5. Managerial Implications

5.1. ROC Analysis and Profitability

Our findings present important insights for decision makers of sports-oriented companies (e.g., retailers or sports clothing brands) and sports organizations. To highlight the impact of our results on the decision-making process, we present a decision tree analysis that shows the key decisions when implementing a targeted advertising approach in Figure 3. The first decision pertains to the level of false alarms the company is willing to allow. We introduced three scenarios in Section 4.2 corresponding to different trade-offs between benefits and cost. Once a scenario has been chosen, the decision maker needs to decide whether to use a predictive model or not. Given the chosen scenario and whether or not the decision maker relies on a predictive model, we can calculate the average profits of the campaign.

To calculate the monetary impact of the three scenarios, we use the equation of Verbraken et al. (2013). They rearranged the formula of Neslin et al. (2006) such that the average profit instead of the total profit is calculated and the false positive and true positive rate are used instead of the top-decile lift. According to Verbraken et al. (2013), the average classification profit, across customers, of a prediction model then becomes

$$Profit_{\text{average}} = B(\gamma(1-\delta)-\phi)TPR - B(\delta+\phi)FPR. \quad (11)$$

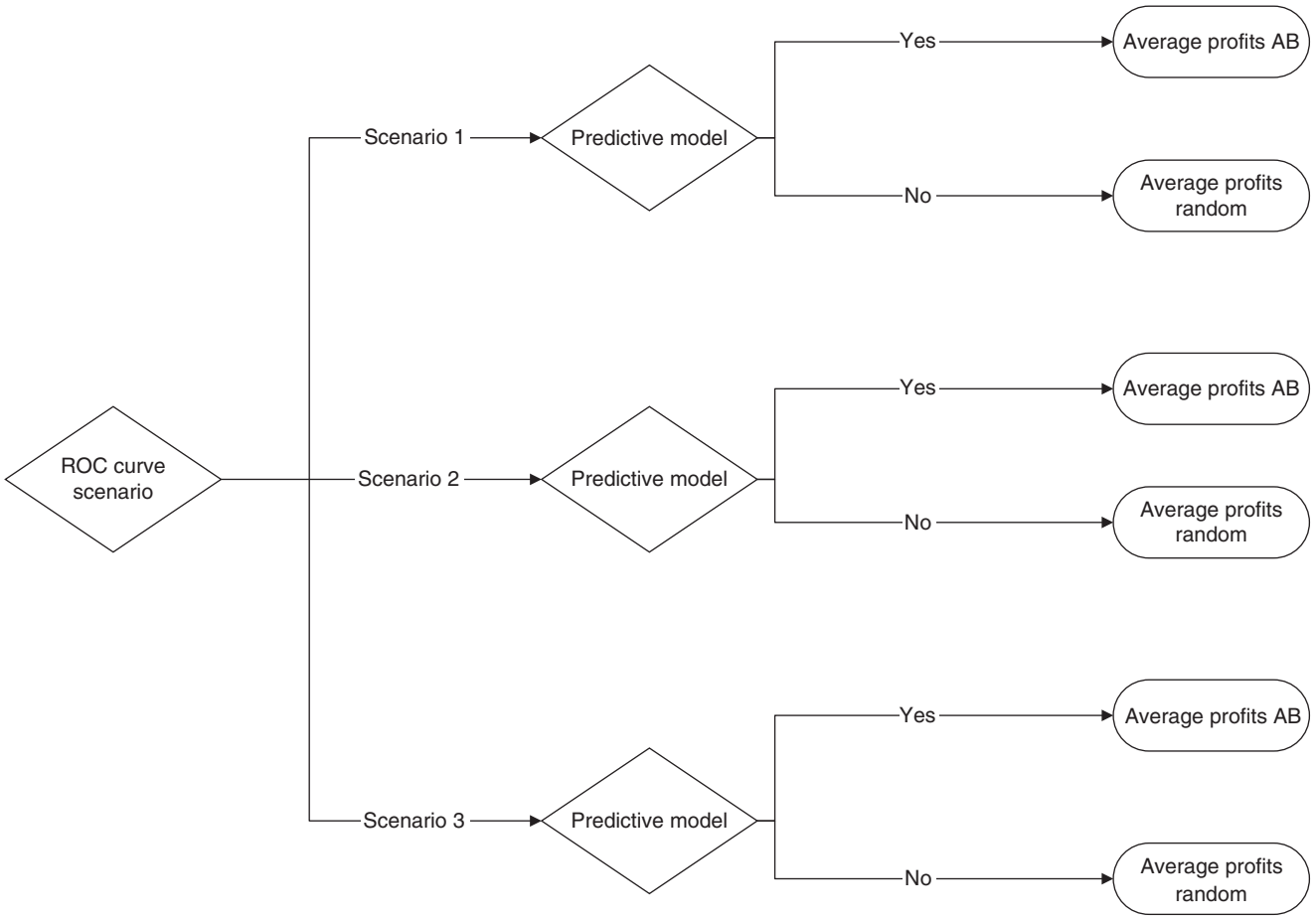
In Equation (11), γ refers to the percentage of identified soccer players that eventually become customers (i.e., the success rate), B is the potential benefits of a future customer, $\delta = d/B$ with d the cost of the incentive provided by the company to persuade the prospect to become a customer and $\phi = f/B$ with f representing the cost related to contacting a user. The parameters γ, δ, ϕ, TPR, and FPR in Equation (11) are dimensionless, and B is expressed in dollars. The first part of Equation (11) represents the total revenues of a campaign, the second part the total costs. We note that the average profit is independent of the targeted customers. Equation (11) informs us that profits will be higher if the benefits of a future customer are higher (B), the campaign is more successful (γ), and the cost of the incentive (d) and contacting a customer (f) are lower. Given a certain threshold, the average profit also increases if the TPR rises or the FPR declines. In contrast, if the FPR goes up or the TPR goes down, the average profit will decline. In that regard Equation (11) is perfectly fit to assess the impact of varying sensitivity and specificity. It allows decision makers to select a certain point on the ROC curve and calculate the impact on average performance.

To provide a flexible decision tool for managers, we consider a wide range of values for all parameters. Table 7 summarizes the range of parameter values and the base case values. We have two base cases: (1) an optimistic situation with a high campaign success rate (γ), high total benefits of the customer (B), and relatively low campaign costs (d) and (2) a conservative situation with low success rate, low customer benefits, and higher campaign costs. The former is denoted as Base(good), the latter as Base(bad). The cost of contacting a customer (f) is set fixed to \$1. We note that both cases represent boundary situations. We do believe that it is important for decision makers to consider all possible scenarios. Given space constraints, however, we do not elaborate on any other possible cases. In the next paragraphs, we perform a detailed sensitivity analysis for the different decisions.

5.2. Sensitivity Analysis

Recall that decision makers are faced with two crucial questions (Figure 3): (1) "Which level of false alarms do we allow?" and (2) "Do we use a predictive model to target our customer?" Based on these decisions, the following scenarios and performance values are used: (1) in scenario 1, the marketer wanted few false positives: $\{FPR = 0.10, TPR_{\text{AB}} = 0.4238, TPR_{\text{Random}} = 0.10\}$,

Figure 3. Decision Tree



where TPR_{AB} is the true positive rate of the adaboost model and TPR_{Random} is the true positive rate of the random model; (2) in scenario 2, the manager accepted a higher number of false positives: $\{FPR = 0.50, TPR_{AB} = 0.9225, TPR_{Random} = 0.50\}$; (3) and in scenario 3, the decision maker allowed a lot of false positives: $\{FPR = 0.90, TPR_{AB} = 0.9974, TPR_{Random} = 0.90\}$. Table 8 summarizes the average profits (per customer) for the two base cases. The parameter values for the base cases can be found in Table 7. We note that for the optimistic base case, all scenarios have a positive average profit. However, for the conservative base case, all scenarios are negative. Even with the adaboost model, the extra benefits of targeting the right customers do not outweigh the costs of a false detection.

Since the average profits for the conservative base case were always negative, we decide to perform a one-way sensitivity analysis for the success rate (γ) and the total benefits (B) in Figure 4. We note that all other parameters are kept constant at the conservative base case values (Table 7). We refer the reader

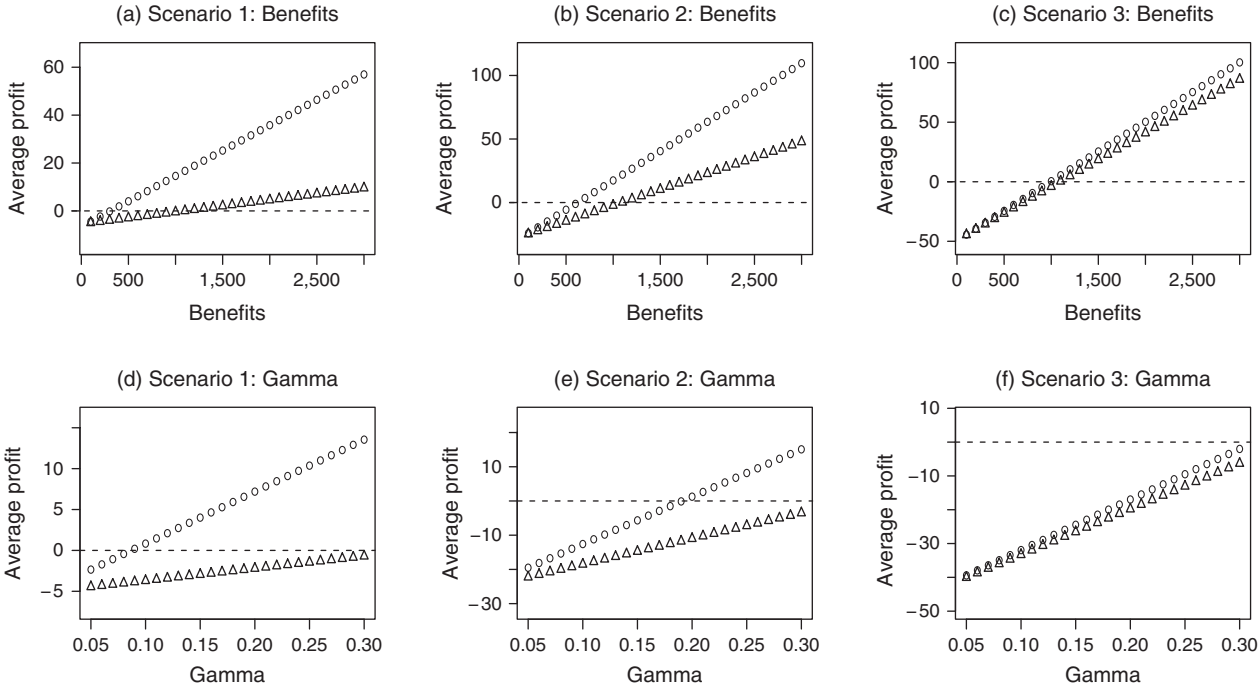
Table 7. Ranges of Parameter Values and Base Cases

Parameters	Min	Max	Base(good)	Base(bad)
B	100	3,000	2,000	200
γ	0.05	0.30	0.20	0.05
d	1	100	20	50

Table 8. Average Profit Simulation for Both Base Cases (in $)

Case	Model	Scenario 1	Scenario 2	Scenario 3
Base(good)	Adaboost	165.30	353.89	375.07
	Random	37.40	187.00	336.60
Base(bad)	Adaboost	−2.35	−19.50	−39.42
	Random	−4.45	−22.25	−40.05

Figure 4. One-Way Sensitivity Analyses for the Benefits and the Success Rate with Conservative Base Values

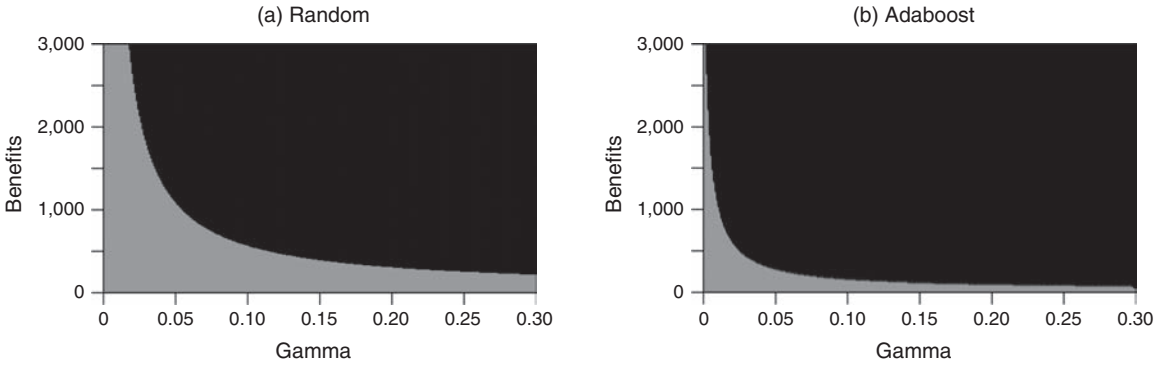


Note. The circles represent the adaboost model, the triangles the random model.

to Appendix D.1 for the one-way sensitivity analyses of companies with optimistic base values. In Figure 4, the circles represent the adaboost model, the triangles the random model. For the total benefits (Figures 4(a)–4(c)), we notice that the difference between the adaboost model and the random case increases when the allowed false alarms decrease (i.e., if the company wants to lower the cost of the campaign). For example, with the adaboost model, the company can already achieve a profit from a customer benefit of \$400 whereas in the random case the benefits have to be at least \$1,000 (Figure 4(a)). The more false positives the company accepts, the smaller the difference between the random model and the adaboost model becomes. However, the adaboost model always achieves higher profits than in the random case. When varying the success rate (Figures 4(d)–4(f)) the difference between the predictive model and random assignment becomes even more clear. We see that in the random case the company is not able to achieve an average profit across all scenarios. However, if the company chooses to implement a predictive model, an average profit occurs if the success rate is 7% in scenario 1 and 17% in scenario 2.

To study the joint effect of the success rate and the total benefits on profitably, we also conducted a two-way sensitivity analysis. Figure 5 plots whether or not the combination of success rate and total benefits yields a profit for the random case (Figure 5(a)) and the adaboost model (Figure 5(b)) for scenario 1. Again, the other parameter values are set to the conservative base case values. The upper right area (black) represents a profit; the lower left (grey) is a loss. We notice that the profit area is much bigger with the adaboost model in comparison to random. For example, if the expected benefits are \$500 and the success rate is 5%, the campaign will not be profitable in the case of random assignment. However, with the same success rate and expected benefits, the adaboost model would yield a profit. We note that the difference between the profit area and loss area decreases when increasing the number of false alarms (e.g., see Appendix D.2). This two-way sensitivity analysis reinforces our finding that a one-to-one advertising campaign can be beneficial for a lot of different companies. Also, we believe that this tool provides managers a good overview of under which conditions they should implement a targeted advertising campaign.

Figure 5. Two-Way Sensitivity Analysis of the Success Rate and Total Benefits for Scenario 1



Note. The grey area represents a loss, the black area a profit.

In sum, we can state that our sensitivity analysis shows there exists a certain trade-off between the allowed false alarms, the true positives, and the average profits. We showed that increasing the false positives is not always beneficial, especially in the case in which the expected benefits and the success rate of the campaign are low. We have also shown the effectiveness of our predictive modeling approach. Especially in the case in which companies want to limit the number of false alarms (scenario 1 and 2), the difference in profitability between our predictive model and the random approach is significant.

6. Conclusion

One of the most important data-sourcing challenges that companies are faced with today manifests itself in the area of prospecting. Where can companies find a large pool of potential customers, and can a sufficient amount of data be captured about these prospects so that a viable decision support system can be developed? The purpose of this study was to develop such a decision support system to help sports brands with the implementation of a targeted advertising approach. We support the following decision: "Which users should we target to sell soccer-related items?" First, we evaluated the feasibility of predicting whether a Facebook user plays soccer or not by applying a broad range of Facebook variables. Second, we benchmarked six algorithms (i.e., logistic regression, random forest, adaboost, kernel factory, neural network, and rotation forest) to determine the best performer. Furthermore, to facilitate the decision analysis process, we included variable importance measures and partial dependence plots. This allows decision makers to gain insight into the most important variables and their relationship with the response. We also conducted a thorough sensitivity analysis on the impact on profitability of implementing our one-to-one strategy. We thereby introduced several scenarios across a wide range of parameter values.

The results clearly indicate that predicting whether a user plays soccer, by applying Facebook data, is a feasible strategy since the AUC ranges from 72.01% to 80.43%, the accuracy from 81.96% to 83.95%, and the top-decile lift from 2.4080 to 3.0550. The best-performing algorithm was adaboost in terms of accuracy and top-decile lift. In terms of AUC, adaboost was the top performer followed by the fusion model (i.e., an aggregator of all the other prediction models), random forest, rotation forest, logistic regression, kernel factory, and neural networks. We note that adaboost, the fusion model, and random forest had an equal statistical performance in terms of AUC. In terms of accuracy and lift, logistic regression did not perform significantly differently in comparison with adaboost and the fusion model. We also included a ROC-curve analysis for the adaboost model and a random model. We elaborated on three possible scenarios for decision makers.

The top-performing variables fall into three categories: general Facebook variables, network variables, and soccer-specific variables. The top-performing variable is the average number of friends who play soccer. The more friends who play soccer, the higher the chances the focal user is a soccer player as well. Hence, we found evidence that users on Facebook tend to choose friends who are alike (McPherson et al. 2001).

There are several other important variables related to fan engagement, such as the *membership in a soccer group*, *number of favorite teams*, *number of favorite teams*, and *number of sports the user likes*. In addition, social network variables were important. The higher users scored on these variables, the higher their engagement in sports, and thus the higher their propensity for playing soccer.

Finally, we believe that our targeted modeling approach has a lot of application fields. For example, sports retailers and sports teams can use this approach to identify and attract new customers. Sports retailers can use it to sell soccer gear, and soccer teams can use it to increase season ticket sales. We conducted a decision tree analysis, which summarizes the most important decisions to implement a targeted advertising campaign. To calculate the impact on profitability, we introduced Equation (11), which allows decision makers to link different points on the ROC curve with the average profits (Verbraken et al. 2013). We employed a wide range of parameter values and two base cases for which we conducted a rigorous sensitivity analysis. Our sensitivity analysis shows that increasing the allowed number of false alarms does not always lead to beneficial results. Especially in the case in which the expected benefits are low, it is utterly important to rely on a predictive model if the company wants to achieve a profit. Therefore, decision makers should carefully assess the trade-off between the benefits and costs of a one-to-one advertising strategy. We believe that our sensitivity analysis can serve as a powerful tool for decision makers to assess the impact of different parameters and scenarios.

In sum, there exist a significant number of application domains, and the predictive potential in each domain is different. This study makes a contribution to literature by making clear recommendations toward companies regarding how to use a data source (i.e., Facebook), encompassing 24% of the world population. By doing so, we solve one of the hardest problems that marketers around the globe are facing: sourcing data for customer acquisition purposes. Specifically, we show that Facebook data can drive accurate models and deliver predictions for very large numbers of consumers. This study lays out clear guidelines in terms of data, algorithms, variables, and sensitivity analysis for advertisers to replicate our system. In addition, we recommend Facebook, Inc., expand its advertising tools to include customized models, such as the ones we develop in this study and make them available alongside their more general targeting options.

7. Limitations and Future Research

Our study is limited because of several reasons. First, for some variables, Facebook only provides the 25 most recent entries. To circumvent this problem, we used the frequency in a specific time period to avoid reaching this limit. We used the last year to calculate the frequency of video uploads and notes, the last four months to create the frequency of check-ins and album uploads. The frequency of status updates, link uploads, and photo uploads were determined by looking at the last seven days.

The second limitation of our study is related to how we gather our data. We extracted our data through a Facebook page of a European soccer team. To gather our data, we made use of a Facebook app that we introduced on the Facebook page of the soccer team. We gave an incentive to Facebook users to subscribe to the app by offering a prize, more specifically, a signed shirt of the soccer team. To avoid privacy issues, we (1) explicitly asked permission from the user, (2) included a section with rules and regulations along with our contact information, (3) promised the user that all information is anonymous and no private messages would be extracted, and finally, (4) added a disclaimer to provide information about the purpose of the research. However, we still acknowledge that part of the Facebook users may not be willing to share their data. For example, this may be because they are not interested in this specific reward or because they are not eager to share personal information. Those who are not willing to share their information may be different from the ones who shared their data. Another downside of our data extraction lies in the fact that we advertised our application through the Facebook page of a European soccer team. Hence, our data set consists of users who were following the team's Facebook page (or had friends who were), and thus, all users are at least slightly interested in soccer. As a result, our sample of Facebook users is not completely representative for the entire population of Facebook. Nevertheless, our goal was to assess in a plausible setting whether identifying soccer players is feasible on Facebook. Companies that want to take this

approach will have access to the same population (some users will donate their data and some will not), and therefore, we cannot say that we have selection effects in our study. We are not generalizing to the entire Facebook population, but we are providing a plausible use case.

Additionally, our research is limited by only including a small number of friends variables, more specifically, the variables that are directly related to sports. Future research could add more friends variables and compare the performance of the algorithms. Interesting variables to include would be the average age of all Facebook friends. As such, practitioners can assess the difference in model performance and determine the most important elements to make accurate predictions.

Another limitation is related to the fact that we only consider self-reported soccer players as our dependent variable. We deliberately chose soccer since it is the biggest sports industry in Europe. However, we believe that a valuable addition to future research may be to broaden the predictions from soccer practitioners to other sports players and see how the results change across the different sports categories.

The final limitation is related to the fact that we only include Facebook data. It might be valuable to augment our data with other types of data. A possible avenue for future research is to add other data coming from several social media platforms. For example, Twitter is a platform that contains a large amount of data, such as tweets, retweets, and likes. Other interesting data platforms are Instagram and Swarm.

Even though this study has several limitations, we are—to the best of our knowledge—the first to evaluate the feasibility of identifying self-reported soccer players using Facebook data. As a result, we believe that this study makes a valuable contribution to literature.

Acknowledgments

We would like to thank the associate editor for the timely management of this manuscript. We also thank the anonymous reviewers and the associate editor for their fruitful comments.

Appendix A. Classification Algorithms

A.1. Regularized Logistic Regression

Logistic regression is a commonly used statistical modeling technique that measures the relationship between a dependent variable and several independent variables by returning probability scores as predicted values for the dependent variable. Logistic regression uses the logistic function, and the estimated relationship is given by the following equation (James et al. 2013, p. 132):

$$p = \frac{1}{1 + e^{-\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_i X_i+\varepsilon}}, \tag{A.1}$$

where p stands for the probability of the outcome of interest, β_0 is the intercept term, $\beta_1, \ldots, \beta_i$ represent the coefficients related to the independent variables $X_1, \ldots, X_i$, and i represents the unique subscript for each independent variable (Tu 1996). The model is fit using the maximum likelihood method (James et al. 2013, p. 130).

When confronted with a lot of independent variables, logistic regression could measure random error or noise and the idiosyncrasies of our data instead of the underlying relationship (i.e., ovefitting) (Babyak 2004). To solve this issue, we use the lasso (least absolute shrinkage and selection operator) approach to regularized logistic regression, which sets a bound on the total sum of the absolute values of all coefficients (Guisan et al. 2002, James et al. 2013).

The shrinkage parameter λ determines how much the coefficients will be shrunk toward zero. Increasing λ will result in smaller coefficients. In the case in which the shrinkage parameter is sufficiently large, it forces some of the coefficients to be exactly zero. As a result, lasso performs variable selection (James et al. 2013, p. 130).

We optimize the lasso shrinkage parameter by means of cross-validation. To fit the model, we use the statistical R-package *glmnet* (Friedman et al. 2015). We set the α parameter to one to obtain the lasso approach and *nlambda* to the default value (100).

A.2. Neural Networks

We use feed-forward neural networks optimized by the BFGS algorithm with one layer of hidden neurons. It has been shown that this optimization technique is more efficient and reliable than backpropagation (Dreiseitl and Ohno-Machado 2002). In advance of using the neural network, we rescale the numerical variables to (−1, 1). We do this by subtracting the midrange from each column, followed by dividing by the half of the range (Ballings et al. 2015). The binary variables remain untouched and are given by the values {0, 1}. Scaling is required to overcome numerical problems, obtain training efficiency, and guard against reaching local optima (Ballings et al. 2015).

We use the statistical *nnet* R-package to implement the algorithm (Ripley and Venables 2015). At the start of the iterative procedure, the initial weights are chosen at random (Ripley 2007, p. 154). Consequently, the results of subsequent neural networks may differ, which can be compared with the human brain (Venkatesh et al. 2014). As recommended by Spackman (1991), we set the *entropy* parameter to the maximum likelihood method. To control the range of initial random weight values, we left the *rang* parameter at the default value (0.5). Additionally, we left the *abstol* parameter

and *rel* parameter to their default values (respectively, 1.0e4 and 1.0e8). To cope with overfitting, we used weight decay (Dreiseitl and Ohno-Machado 2002). The maximum number of iterations (*maxit*) and the maximum number of weights (*MaxNWts*) both equal 5,000 as to avoid the possibility of early stopping. Finally, the number of nodes in the hidden layer and the weight decay factor are determined by performing a grid search (Dreiseitl and Ohno-Machado 2002). The optimal combination was selected by sequencing over all combinations of *decay* = {0.001, 0.01, 0.1} (Ripley 2007, p. 163) and *size* = [1, 2, . . . , 20] (Ripley 2007, p. 170).

A.3. Random Forest

Random forest is an ensemble technique that copes with the suboptimal performance of decision trees (lack of robustness) (Dudoit et al. 2002). Random forest uses a combination of tree predictors in which each tree depends on the values of a random vector. To build these ensembles, bagging is used to grow each tree (Breiman 2001). Bagging (or bootstrap aggregation) creates new training data sets by randomly sampling with replacement from the original data set. Bagging improves classification accuracy by lowering the variance of classification errors, and thus, random forest copes with the instability of a classifier (Breiman 2001, Chan and Paelinckx 2008, Gislason et al. 2006).

In combination with bagging, random forest uses random feature selection to construct multiple trees. This implies that at each node split only a random subset of the predictors is considered. The final ensemble is built by means of majority voting across all bootstrapped trees (Breiman 2001).

Random forest only requires two parameters: the number of trees and the number of variables to try at each split. Along with Breiman's recommendation, we use a large number of trees (500), and we set the number of variables equal to the square root of the total number of variables. To create our model, we use the *randomForest* R-package (Liaw and Wiener 2002).

A.4. Adaboost

In the initial adaboost algorithm, every model is built sequentially by reweighting the training data (James et al. 2013, p. 322). At each iteration, incorrectly classified instances receive more weight than correctly classified observations. As such the model focusses on instances that are hard to classify (Chan and Paelinckx 2008, Freund et al. 1996). The final prediction model is the weighted sum of the previous models (Chan and Paelinckx 2008, James et al. 2013, Freund et al. 1996).

We use a recent variant of the initial adaboost algorithm, namely stochastic boosting, which includes randomness as an integral part of the procedure by drawing bootstrap samples at every iteration. The chance that an observation is selected is proportional to the weight in the current iteration (Friedman 2002).

Two important parameters have to be set: the number of iterations and the number of terminal nodes. In line with the recommendation of Friedman (2002), we set the number of iterations to 500. In addition, we set the maximum depth of our trees to three to determine the number of terminal nodes. We use the R-package *ada* to fit our stochastic boosting model (Culp et al. 2006).

A.5. Kernel Factory

Kernel factory is built by randomly dividing the data into row and column partitions (Ballings and Van den Poel 2013). Each partition is mutually exclusive and transformed into a kernel matrix K by a kernel function. The *burn* method is used to automatically select the best kernel function (polynomial, radial base, or linear). In a next phase, a random forest model is built for each kernel matrix K. This leads to a number of predictions equivalent to the number of mutually exclusive partitions. The final predictions are calculated by taking the weighted averages optimized by a genetic algorithm (Ballings and Van den Poel 2013).

The kernel factory approach has the advantage of inducing both diversity and accuracy. While diversity is augmented because the partitions are based on randomly selected features and observations, accuracy is preserved by the kernel function and genetic algorithm (Ballings and Van den Poel 2013). Kernel factory is implemented with the statistical R-package *kernelFactory* of Ballings and Van den Poel (2015b).

A.6. Rotation Forest

Rodriguez et al. (2006) propose an ensemble based on feature selection. The rotation matrix is used to create the training data that can be used by a base classifier. The rotation matrix is created as follows (Rodriguez et al. 2006). First, the feature set is split into K subsets. Second, a bootstrap sample is built for every subset followed by applying principal component analysis (PCA). The goal of PCA is to create a low-dimensional representation of the data by finding an orthogonal transformation of the variables, called principal components. Every dimension found by PCA is a linear combination of the features (De Bock and Van den Poel 2011). The coefficients obtained by PCA are stored in a matrix. Next, the rotation matrix is created by rearranging the matrix in the order that matches the original features. The final step in building the training set is to transform the original training set using the rotation matrix (De Bock and Van den Poel 2011). Predictions are made by applying decision trees as base classifier on every new training set. Decision trees are used because of their sensitivity to rotation of the feature axes (Rodriguez et al. 2006). Rotation forest builds an ensemble that is both diverse and accurate. Diversity is induced by the feature selection for every base classifier while accuracy is promoted by keeping all principal components and using the entire data set to train every base classifier (Rodriguez et al. 2006).

We use the statistical R-package *rotationForest* of Ballings and Van den Poel (2015b) to implement rotation forest. The number of base classifiers (L) and the number of variable subsets (K) were both set to their default (respectively, 10 and three) (Ballings and Van den Poel 2015b).

Appendix B. Information-Fusion Sensitivity Analysis

Table B.1. Top 35 Variables Based on Information Fusion

Rank	Variable name	Sensitivity score	Type
1	*AVERAGE*(*friends playing soccer*)	0.0204	*N*
2	*IND*(*soccer group*)	0.0189	*S*
3	*COUNT*(*favorite teams*)	0.0162	*S*
4	*IND*(*gender == female*)	0.0155	*G*
5	*IND*(*"interested in" present*)	0.0090	*G*
6	*IND*(*relationship status == single*)	0.0070	*G*
7	*Age*	0.0063	*G*
8	*COUNT*(*likes related to soccer*)	0.0061	*S*
9	*IND*(*biography present*)	0.0059	*G*
10	*IND*(*education == college*)	0.0053	*G*
11	*COUNT*(*likes sports category*)	0.0052	*S*
12	*COUNT*(*television shows*)	0.0052	*G*
13	*IND*(*relationship status == relationship*)	0.0046	*G*
14	*IND*(*relationship status == married*)	0.0041	*G*
15	*COUNT*(*friends*)	0.0038	*G*
16	*COUNT*(*music*)	0.0036	*G*
17	*IND*(*relationship status present*)	0.0035	*G*
18	*IND*(*quotes present*)	0.0034	*G*
19	*COUNT*(*movies*)	0.0034	*G*
20	*COUNT*(*interest == soccer*)	0.0034	*S*
21	*RECENCY*(*album comments*)	0.0032	*G*
22	*COUNT*(*sports other than soccer*)	0.0030	*S*
23	*IND*(*like TV show*)	0.0029	*G*
24	*IND*(*religion present*)	0.0029	*G*
25	*AVERAGE*(*albums for friends*)	0.0028	*G*
26	*COUNT*(*languages*)	0.0024	*G*
27	*IND*(*education == high school*)	0.0024	*G*
28	*AVERAGE*(*soccer groups over all friends*)	0.0023	*N*
29	*COUNT*(*likes music category*)	0.0022	*G*
30	*IND*(*hometown present*)	0.0021	*G*
31	*COUNT*(*likes related soccer*)	0.0019	*S*
32	*AVERAGE*(*public albums*)	0.0019	*G*
33	*COUNT*(*likes*)	0.0017	*G*
34	*COUNT*(*TV show related to soccer*)	0.0017	*S*
35	*RECENCY* (*video tags*)	0.0017	*G*

Appendix C. Partial Dependence Plots

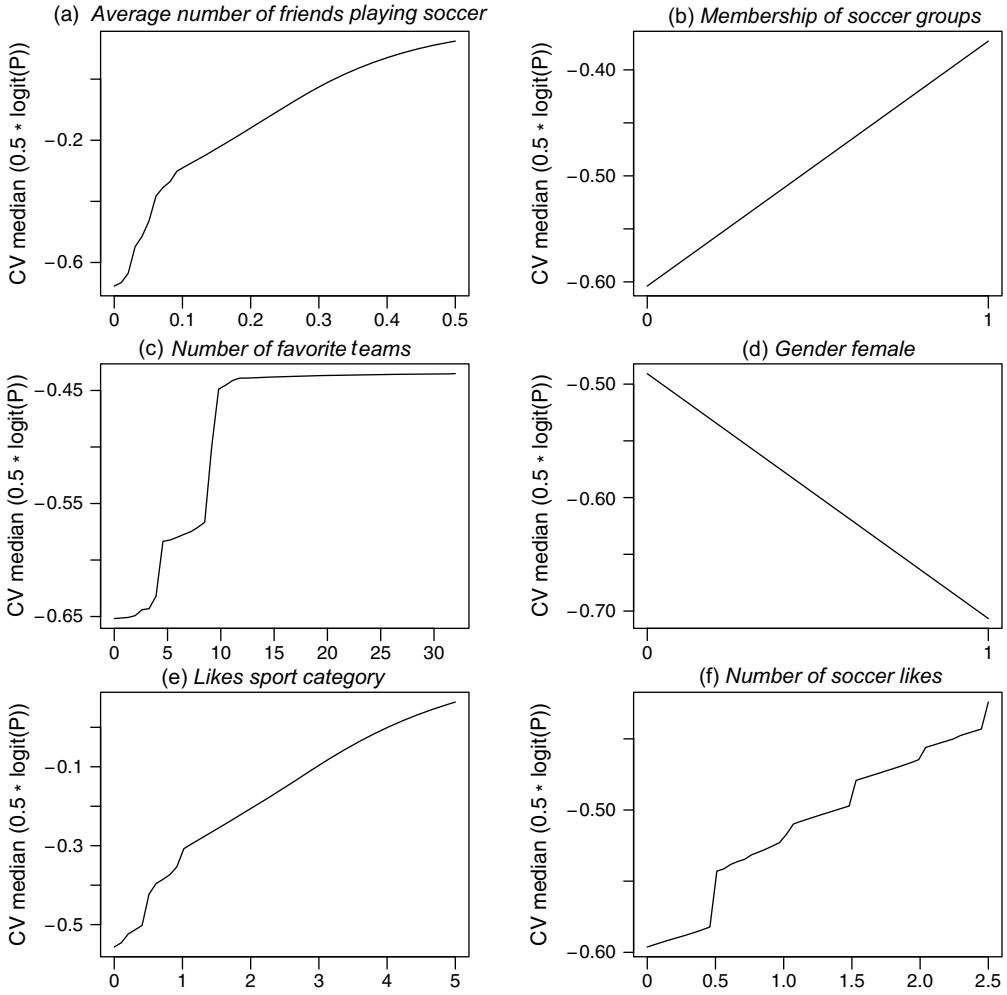
We visualize the relationship between the most important variables and the response by utilizing partial dependence plots (PDP). PDP allow us to consider the relationship while eliminating the effect of the other independent variables (Friedman and Meulman 2003). We follow the method of Berk (2008) to create partial dependence plots. First, a fusion model based on Equation (9) is built on the original sample. Next, for every distinct value v of a variable x, a new data set is built that only takes on that one value v while leaving all other variables untouched. Next, we predict the response for every new data set using the ensemble model that has been created on the original data set. This is followed by taking the mean of half the logit of the predictions, resulting in one single value p for all instances. Finally, we plot all values v against their corresponding p (Berk 2008). In Figure C.1, we depict the relationship between the response and the top six variables.

Table 6 clearly indicates that the *average number of someone's Facebook friends playing soccer* is the most important variable. Figure C.1(a) illustrates that the proportion of friends who play soccer is positively related with the dependent variable. The more friends who play soccer, the more points of interaction the user will have with soccer and the more likely that the user will be interested in playing soccer. A possible explanation in extant literature is provided by the principle of homophily in social networks (McPherson et al. 2001). The theory states that people spend time with people who share their personal characteristics, behavioral features, and socio-demographic characteristics. As such, homophily limits users' social world by the information they are exposed to, the attitudes they form, and the interactions they have (McPherson et al. 2001). This is also in line with Bogaert et al. (2016a), who found that the number of friends who attend an event have a positive influence on the propensity of attending the focal event.

Another important variable is the *number of favorite teams*. Figure C.1(c) indicates a strong positive relationship between the number of favorite teams of a Facebook user and our dependent variable. We found similar relationships for the variables that indicate *whether or not the user is part of a soccer group on Facebook* (Figure C.1(b)), the *number of soccer likes* (Figure C.1(f)), and the *number of likes related to a sport category* (Figure C.1(e)). The relationship of the number of favorite teams, soccer groups, soccer-related likes, and likes related to a sport category can be explained by looking at theories concerning fan loyalty and fan engagement. Several studies demonstrate the importance of fan loyalty and engagement in the purchase intentions of the customer (Bauer et al. 2008, Funk 1998, Yoshida et al. 2014). The *number of favorite teams*, *membership in a soccer group*, *number of soccer likes*, and *liking a sport category* are all indicators of users having a positive attitude toward soccer. Therefore, we can consider them as variables that indicate the degree of engagement or loyalty (Funk 1998, Yoshida et al. 2014, Bauer et al. 2008). The higher the degree of engagement for a certain sport, the higher the probability that the user is also practicing the focal sport. Consequently, a higher value on these variables indicates a higher engagement toward soccer and thus corresponds to a higher probability of playing soccer.

Finally, we see that *whether or not the user is male or female* has an influence on the chances of playing soccer (Figure C.1(d)). The chances of playing soccer are smaller when the user is female.

Figure C.1. Partial Dependence Plots (PDP)



Appendix D. Sensitivity Analysis

D.1. One-Way Sensitivity Analysis

Figure D.1 plots the one-way sensitivity analysis for the success rate and the benefits of a campaign with optimistic base values (see Table 7). We notice that with optimistic parameter values the adaboost model always yields a profit while the random model is still too costly with small expected benefits.

D.2. Two-Way Sensitivity Analysis

Figure D.2 depicts the trade-off between the success rate of a campaign and the expected benefits of a customer for scenario 3. We see that the difference between the profit area (black) and the loss area (grey) are almost the same for the adaboost model and the random model. This was expected since the differences between the true positives rates are

Figure D.1. One-Way Sensitivity Analyses for the Benefits and the Success Rate with Optimistic Base Values

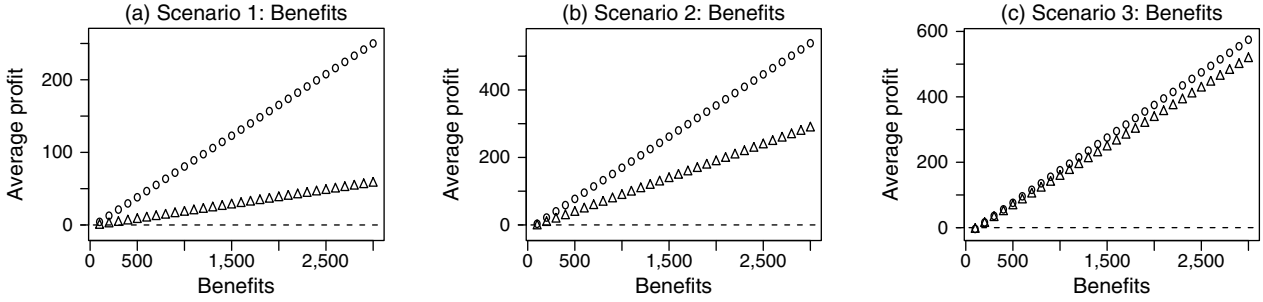
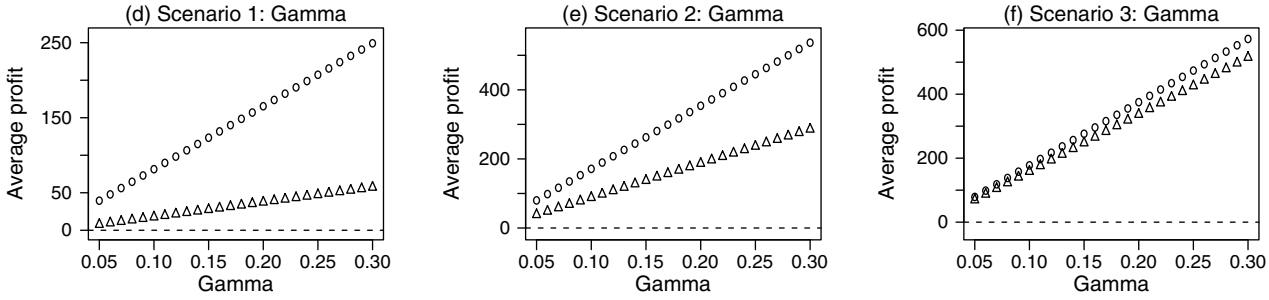
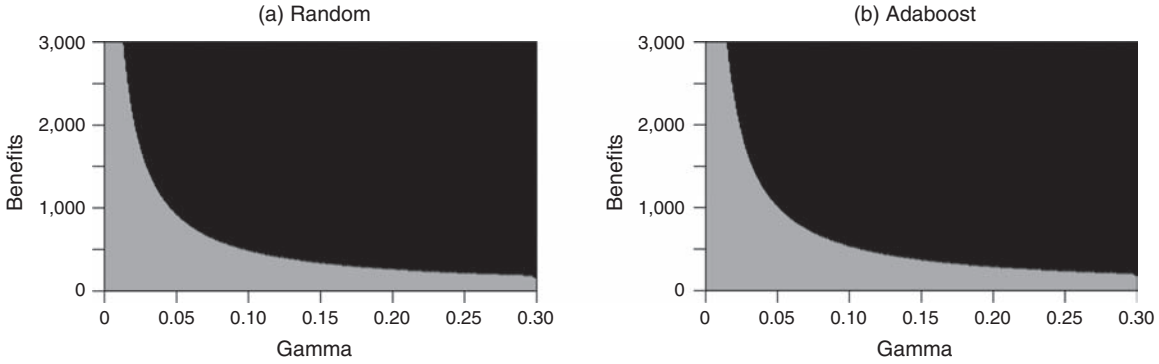


Figure D.1. (Continued)



Note. The circles represent the adaboost model, the triangles the random model.

Figure D.2. Two-Way Sensitivity Analysis of the Success Rate and Total Benefits for Scenario 3



Note. The grey area represents a loss, the black area a profit.

rather small between the two models in scenario 3 (see Figure 1); this can also be seen in the one-way sensitivity analysis of both parameters (see Figures 4(c) and 4(f)). In general, we can state that the more false alarms a company accepts, the smaller the difference between the adaboost model and the random case.

Endnote

[1] We use the term "Facebook, Inc." to indicate the company whereas the term "Facebook" refers to the social media platform.

References

Baatarjav EA, Phithakkitnukoon S, Dantu R (2008) Group recommendation system for facebook. Meersman R, Tari Z, Herrero P, eds. *On the Move to Meaningful Internet Systems*: *OTM* 2008 *Workshops* (Springer, Berlin), 211–219.

Babyak MA (2004) What you see may not be what you get: A brief, nontechnical introduction to overfitting in regression-type models. *Psychosomatic Medicine* 66(3):411–421.

Baldi P, Hornik K (1989) Neural networks and principal component analysis: Learning from examples without local minima. *Neural Networks* 2(1):53–58.

Ballings M, Van den Poel D (2012) Customer event history for churn prediction: How long is long enough? *Expert Systems Appl.* 39(18):13517–13522.

Ballings M, Van den Poel D (2013) Kernel factory: An ensemble of kernel machines. *Expert Systems Appl.* 40(8):2904–2913.

Ballings M, Van den Poel D (2015a) CRM in social media: Predicting increases in Facebook usage frequency. *Eur. J. Oper. Res.* 244(1): 248–260.

Ballings M, Van den Poel D (2015b) R-package *rotationForest*: Fit and deploy rotation forest models, v. 0.1.3. Retrieved, https://CRAN.R-project.org/package=rotationForest.

Ballings M, Van den Poel D, Hespeels N, Gryp R (2015) Evaluating multiple classifiers for stock price direction prediction. *Expert Systems Appl.* 42(20):7046–7056.

Bauer HH, Stokburger-Sauer NE, Exler S (2008) Brand image and fan loyalty in professional team sport: A refined model and empirical assessment. *J. Sport Management* 22(2):205–226.

Belzer J (2016) Sports industry 101: Breaking into the business of sports. *Forbes* (February 5), http://www.forbes.com/sites/jasonbelzer/2014/02/05/sports-industry-101-breaking-into-the-business-of-sports/.

Berjani B, Strufe T (2011) A recommendation system for spots in location-based online social networks. *Proc. 4th Workshop Soc. Network Systems* (Association for Computing Machinery, New York), 1–6.

Berk RA (2008) *Statistical Learning from a Regression Perspective* (Springer Science and Business Media, Berlin).

Bogaert M, Ballings M, Van den Poel D (2016a) The added value of Facebook friends data in event attendance prediction. *Decision Support Systems* 82(February):26–34.

Bogaert M, Ballings M, Van den Poel D (2016b) Evaluating the importance of different communication types in romantic tie prediction on social media. *Ann. Oper. Res.*, ePub ahead of print August 17, https://dx.doi.org/10.1007/s10479-016-2295-0.

Breiman L (2001) Random forests. *Machine Learn.* 45(1):5–32.

Bu J, Tan S, Chen C, Wang C, Wu H, Zhang L, He X (2010) Music recommendation by unified hypergraph: Combining social media information and music content. del Bimbo A, Chang S-F, Smeulders A, eds. *Proc. 18th ACM Internat. Conf. Multimedia* (Association for Computing Machinery, New York), 391–400.

Burez J, Van den Poel D (2007) CRM at a pay-TV company: Using analytical models to reduce customer attrition by targeted marketing for subscription services. *Expert Systems Appl.* 32(2):277–288.

Carmagnola F, Vernero F, Grillo P (2009) Sonars: A social networks-based algorithm for social recommender systems. Houben G-J, McCalla G, Pianesi F, Zancanaro M, eds. *User Modeling, Adaptation, Personalization* (Springer, Berlin), 223–234.

Cesario E, Congedo C, Marozzo F, Riotta G, Spada A, Talia D, Trunfio P, Turri C (2015) Following soccer fans from geotagged tweets at FIFA World Cup 2014. Zhou M, Lin J, Wu K, Fang L, eds. 2015 2*nd IEEE Internat. Conf. Spatial Data Mining and Geographical Knowledge Services* (*ICSDM*) (IEEE, Hoboken, NJ), 33–38.

Cesario E, Iannazzo AR, Marozzo F, Morello F, Riotta G, Spada A, Talia D, Trunfio P (2016) Analyzing social media data to discover mobility patterns at EXPO 2015: Methodology and results. Zeljkovic V, ed. 2016 *Internat. Conf. High Performance Comput. Simulation* (*HPCS*) (IEEE, Hoboken, NJ), 230–237.

Chan JCW, Paelinckx D (2008) Evaluation of random forest and adaboost tree-based ensemble classification and spectral band selection for ecotope mapping using airborne hyperspectral imagery. *Remote Sensing Environ.* 112(6):2999–3011.

Chang J, Sun E (2011) Location 3: How users share and respond to location-based data on social networking sites. Nicolov N, Shanahan JG, eds. *Proc. Fifth Internat. AAAI Conf. Weblogs Soc. Media* (Association for the Advancement of Artificial Intelligence, Menlo Park, CA), 74–80.

Coussement K, Benoit DF, Van den Poel D (2010) Improved marketing decision making in a customer churn prediction context using generalized additive models. *Expert Systems Appl.* 37(3): 2132–2143.

Culp M, Johnson K, Michailidis G (2006) ada: An *r* package for stochastic boosting. *J. Statist. Software* 17(2):1–27.

De Bock KW, Van den Poel D (2011) An empirical evaluation of rotation-based ensemble classifiers for customer churn prediction. *Expert Systems Appl.* 38(10):12293–12301.

Deloitte (2016) Deloitte football money league 2017. Accessed October 6, 2016, http://www2.deloitte.com/uk/en/pages/sports-business-group/articles/deloitte-football-money-league.html.

Demšar J (2006) Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.* 7(December):1–30.

D'Haen J, Van den Poel D, Thorleuchter D, Benoit DF (2016) Integrating expert knowledge and multilingual web crawling data in a lead qualification system. *Decision Support Systems* 82(February):69–78.

Dietterich TG (1998) Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Comput.* 10(7):1895–1923.

Dietterich TG (2000) Ensemble methods in machine learning. Kittler J, Roli F, eds. *Multiple Classifier Systems* (Springer, Berlin), 1–15.

Dreiseitl S, Ohno-Machado L (2002) Logistic regression and artificial neural network classification models: A methodology review. *J. Biomedical Informatics* 35(5):352–359.

Dudoit S, Fridlyand J, Speed TP (2002) Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Amer. Statist. Assoc.* 97(457):77–87.

Egan K (2016) The difference between Facebook, Twitter, Linkedin, Google+, YouTube, and Pinterest. (May 19), https://www.impactbnd.com/blog/the-difference-between-facebook-twitter-linkedin-google-youtube-pinterest.

Facebook (2017) Company info Facebook newsroom. Accessed January 27, 2017, http://newsroom.fb.com/company-info/.

Filo K, Lock D, Karg A (2015) Sport and social media research: A review. *Sport Management Rev.* 18(2):166–181.

Freund Y, Schapire RE, others (1996) Experiments with a new boosting algorithm. *Internat. Conf. Machine Learning* 96(July):148–156.

Friedman J, Hastie T, Simon N, Tibshirani R (2015) R-package *glmnet*: Lasso and elastic-net regularized generalized linear models, v.2.0-13. Retrieved, https://CRAN.R-project.org/package=glmnet.

Friedman JH (2002) Stochastic gradient boosting. *Comput. Statist. Data Anal.* 38(4):367–378.

Friedman JH, Meulman JJ (2003) Multiple additive regression trees with application in epidemiology. *Statist. Medicine* 22(9): 1365–1381.

Funk DC (1998) Fan loyalty: The structure and stability of an individual's loyalty toward an athletic team. Doctoral dissertation, Ohio State University, Columbus.

Gao H, Tang J, Hu X, Liu H (2013) Exploring temporal effects for location recommendation on location-based social networks. Yang Q, King I, Li Q, eds. *Proc. 7th ACM Conf. Recommender Systems* (Association for Computing Machinery, New York), 93–100.

Gislason PO, Benediktsson JA, Sveinsson JR (2006) Random forests for land cover classification. *Pattern Recognition Lett.* 27(4): 294–300.

Golbeck J, Hansen D (2011) Computing political preference among Twitter followers. Tan D, ed. *Proc. SIGCHI Conf. Human Factors Comput. Systems* (Association for Computing Machinery, New York), 1105–1108.

Guisan A, Edwards TC Jr, Hastie T (2002) Generalized linear and generalized additive models in studies of species distributions: Setting the scene. *Ecological Modelling* 157(2–3):89–100.

Guy I, Zwerdling N, Ronen I, Carmel D, Uziel E (2010) Social media recommendation based on people and tags. Crestani F, Marchand-Maillet S, eds. *Proc. 33rd Internat. ACM SIGIR Conf. Res. Development Inform. Retrieval* (Association for Computing Machinery, New York), 194–201.

Hanley JA, McNeil BJ (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 143(1):29–36.

He H, Garcia EA (2009) Learning from imbalanced data. *IEEE Trans. Knowledge and Data Engrg.* 21(9):1263–1284.

Hernandez-Orallo J, Flach P, Ferri C (2012) A unified view of performance metrics: Translating threshold choice into expected classification loss. *J. Machine Learn. Res.* 13(October):2813–2869.

Hoch SJ (1988) Who do we know: Predicting the interests and opinions of the american consumer. *J. Consumer Res.* 15(3):315–324.

Hsu CC, Chen HC, Huang KK, Huang YM (2012) A personalized auxiliary material recommendation system based on learning style on Facebook applying an artificial bee colony algorithm. *Comput. Math. Appl.* 64(5):1506–1513.

James G, Witten D, Hastie T, Tibshirani R (2013) *An Introduction to Statistical Learning*, Vol. 112 (Springer, Berlin).

Janitza S, Strobl C, Boulesteix AL (2013) An AUC-based permutation variable importance measure for random forests. *BMC Bioinformatics* 14(1):1–11.

Kalampokis E, Tarabanis K, Tambouris E (2013) Understanding the predictive power of social media. *Internet Res.* 23(5):544–559.

Kayaalp M, Özyer T, Özyer ST (2009) A collaborative and content based event recommendation system integrated with data

collection scrapers and services at a social networking site. Harkiolakis N, ed. *Proc. 2009 Internat. Conf. Advances Soc. Network Anal. Mining* (IEEE Computer Society, Washington, DC), 113–118.
Kim HN, Saddik AE (2013) Exploring social tagging for personalized community recommendations. *User Model User-Adap. Inter.* 23(2–3):249–285.
Kisilevich S, Keim D, Rokach L (2010) A novel approach to mining travel sequences using collections of geotagged photos. Painho M, Santos MY, Pundt H, eds. *Geospatial Thinking* (Springer, Berlin), 163–182.
Konstas I, Stathopoulos V, Jose JM (2009) On social networks and collaborative recommendation. Allan J, Aslam J, eds. *Proc. 32nd Internat. ACM SIGIR Conf. Res. Development Inform. Retrieval* (Association for Computing Machinery, New York), 195–202.
Lampe CA, Ellison N, Steinfield C (2007) A familiar Face(Book): Profile elements as signals in an online social network. Rosson MB, ed. *Proc. SIGCHI Conf. Human Factors Comput. Systems* (Association for Computing Machinery, New York), 435–444.
Langley P (2000) Crafting papers on machine learning. *Internat. Conf. Machine Learning* (*ICML*), 1207–1216.
Lemmens A, Croux C (2006) Bagging and boosting classification trees to predict churn. *J. Marketing Res.* 43(2):276–286.
Liaw A, Wiener M (2002) Classification and regression by randomForest. *R news* 2(3):18–22.
Malthouse EC, Haenlein M, Skiera B, Wege E, Zhang M (2013) Managing customer relationships in the social media era: Introducing the social CRM house. *J. Interactive Marketing* 27(4):270–280.
McPherson M, Smith-Lovin L, Cook JM (2001) Birds of a feather: Homophily in social networks. *Annual Rev. Sociol.* 27(1):415–444.
Mesnage CS, Rafiq A, Dixon S, Brixtel RP (2011) Music discovery with social networks. Anglade A, Celma O, Fields B, Lamere P, McFee B, eds. *Workshop on Music Recommendation and Discovery* (Association for Computing Machinery, New York), 1–6.
Neslin SA, Gupta S, Kamakura W, Lu J, Mason CH (2006) Defection detection: Measuring and understanding the predictive accuracy of customer churn models. *J. Marketing Res.* 43(2):204–211.
Oztekin A, Delen D, Turkyilmaz A, Zaim S (2013) A machine learning-based usability evaluation method for eLearning systems. *Decision Support Systems* 56(December):63–73.
Oztekin A, Kizilaslan R, Freund S, Iseri A (2016) A data analytic approach to forecasting daily stock returns in an emerging market. *Eur. J. Oper. Res.* 253(3):697–710.
Passant A, Raimond Y (2008) Combining social music and semantic web for music-related recommender systems. Finin T, ed. *7th Internat. Semantic Web Conf. Citeseer* (Semantic Web Science Association, Karlsruhe, Germany), 1–6.
Predd JB, Osherson DN, Kulkarni SR, Poor HV (2008) Aggregating probabilistic forecasts from incoherent and abstaining experts. *Decision Anal.* 5(4):177–189.
Quijano-Sanchez L, Recio-Garcia JA, Diaz-Agudo B (2011) Happymovie: A Facebook application for recommending movies to groups. Grégoire E, ed. *2011 IEEE 23rd Internat. Conf. Tools Artificial Intelligence* (IEEE, Hoboken, NJ), 239–244.
Ripley B, Venables W (2015) R-package *nnet*: Feed-forward neural networks and multinomial log-linear models, v. 7.3-12. Retrieved, https://CRAN.R-project.org/package=nnet.
Ripley BD (2007) *Pattern Recognition and Neural Networks* (Cambridge University Press, Cambridge, UK).
Rodriguez JJ, Kuncheva LI, Alonso CJ (2006) Rotation forest: A new classifier ensemble method. *Pattern Anal. Machine Intelligence, IEEE Trans.* 28(10):1619–1630.
Rosaci D, Sarne GML (2014) Matching users with groups in social networks. Zavoral F, Jung JJ, Badica C, eds. *Intelligent Distributed Computing*, Vol. 7 (Springer International Publishing, Berlin), 45–54.
Said A, De Luca EW, Albayrak S (2011) Using social and pseudosocial networks for improved recommendation quality. *IJCAI 2011—9th Workshop Intelligent Techniques Web*, 45–48.
Sevim C, Oztekin A, Bali O, Gumus S, Guresen E (2014) Developing an early warning system to predict currency crises. *Eur. J. Oper. Res.* 237(3):1095–1104.
Shapira B, Rokach L, Freilikhman S (2012) Facebook single and cross domain data for recommendation systems. *User Model User-Adapted Interaction* 23(2–3):211–247.
Spackman KA (1991) Maximum likelihood training of connectionist models: Comparison with least squares back-propagation and logistic regression. Clayton PD, ed. *Proc. Annual Sympos. Comput. Appl. Medical Care* (American Medical Informatics Association, Bethesda, MD), 285–289.
Specht DF (1990) Probabilistic neural networks. *Neural Networks* 3(1): 109–118.
Tu JV (1996) Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *J. Clinical Epidemiology* 49(11):1225–1231.
Tumasjan A, Sprenger TO, Sandner PG, Welpe IM (2010) Predicting elections with twitter: What 140 characters reveal about political sentiment. Hears M, ed. *4th Internat. Conf. Weblogs Social Media* (*ICWSM*), Vol. 10 (Association for the Advancement of Artificial Intelligence, Menlo Park, CA), 178–185.
Ulvila JW, Gaffney JE Jr (2004) A decision analysis method for evaluating computer intrusion detection systems. *Decision Anal.* 1(1):35–50.
Venkatesh K, Ravi V, Prinzie A, Van den Poel D (2014) Cash demand forecasting in ATMs by clustering and neural networks. *Eur. J. Oper. Res.* 232(2):383–392.
Verbraken T, Verbeke W, Baesens B (2013) A novel profit maximizing metric for measuring classification performance of customer churn prediction models. *IEEE Trans. Knowledge Data Engrg.* 25(5):961–973.
Wang G, Kulkarni SR, Poor HV, Osherson DN (2011) Aggregating large sets of probabilistic forecasts by weighted coherent adjustment. *Decision Anal.* 8(2):128–144.
Wei P, Lu Z, Song J (2015) Variable importance analysis: A comprehensive review. *Reliability Engrg. System Safety* 142:399–432.
Yoshida M, Gordon B, Nakazawa M, Biscaia R, others (2014) Conceptualization and measurement of fan engagement: Empirical evidence from a professional sport context. *J. Sport Management* 28(4):399–417.
Zhang Y, Wu H, Sorathia V, Prasanna VK (2013) Event recommendation in social networks with linked data enablement. Hammoudi S, Maciaszek L, Cordeiro J, Dietz J, eds. *Proc. 15th Internat. Conf. Enterprise Inform. Systems* (*ICEIS*), Vol. 2 (Springer, Berlin), 371–379.

Matthias Bogaert is a Ph.D. researcher and teaching assistant at Ghent University. He received his B.S. in business engineering and M.S. in business engineering: marketing engineering/data analytics from Ghent University. His research interests are social media analytics, predictive analytics, CRM, and machine learning. He has published in journals such as *Annals of Operational Research*, *Decision Support Systems*, and *Omega*.

Michel Ballings is an assistant professor in the Department of Business Analytics and Statistics at the University of Tennessee (Knoxville). He teaches data mining, customer analytics, and social media and web analytics. His research focuses on the intersection of computing, machine learning,

and business. He has published in journals, such as *European Journal of Operational Research*, *Omega*, *Decision Support Systems*, *Journal of Product Innovation Management*, and *Expert Systems with Applications*.

Martijn Hosten is a masters student in business engineering at Ghent University. He has received his B.S. in business engineering and M.S. in business engineering: data analytics from Ghent University. His research interests include social media analytics.

Dirk Van den Poel is a professor of data analytics/big data at Ghent University, Belgium. He teaches courses such as big data, Apache Spark, analytical customer relationship management, advanced predictive analytics, and predictive and prescriptive analytics. His major research interests are in the field of analytical CRM: customer acquisition, churn, upsell/cross-sell, and win-back modeling. His methodological interests include ensemble classification methods and big data analytics.